AIX-MARSEILLE UNIVERSITE

*****

AIX MARSEILLE SCHOOL OF ECONOMICS (AMSE)

*****

SECRETARIAT GENERAL DES AFFAIRES REGIONALES (SGAR)

RAPPORT D'ALTERNANCE

France 2030 : Aller vers les entreprises innovantes

Master 2 Econométrie Big Data et Statistique / Magistère Ingénieur Economiste

Etudiant :
M. Ulrich Gbènakpon SEGODO

Tuteur Académique :
M. Emmanuel FLACHAIRE

Tuteur Professionel :
Mme. Fenitra DUPONT-RAZANAJATOVO

23 août 2024

Engagement de Non-Plagiat

Je soussigné, Ulrich SEGODO, déclare être pleinement consciente que le plagiat par copie de documents ou d'une partie d'un document publiée sous toutes ses formes et sur tous les supports, y compris les publications en ligne, constitue une violation des droits d'auteur et des droits voisins, ainsi qu'une fraude pure et simple.

En conséquence, je m'engage à citer toutes les sources et tous les auteurs que j'ai utilisés pour rédiger mon rapport et ses annexes.

Date : 23 août 2024

Signature

Ulrich SEGODO

Remerciements

Je tiens à exprimer mes sincères remerciements à M. Tanguy VAN YPERSELE, directeur de l'école AMSE ainsi qu'à M. Badih GHATTAS et M. Christian SCHULTER, les responsables pédagogiques du Master 2 EBDS. J'aimerais également remercier M. Emmanuel FLACHAIRE pour son accompagnement et ses précieux conseils au cours de la rédaction de ce rapport.

Mes remerciements vont également à M. Didier Mamis, Secrétaire Général du SGAR, et à M. Olivier TEISSIER, secrétaire adjoint pôle politiques publiques, pour m'avoir accueillie au sein de leur équipe. Je tiens tout particulièrement à remercier ma tutrice professionnelle, Mme. Fenitra DUPONT-RAZANAJATOVO dont l'expertise en tant que chargée de mission numérique a été inestimable. Elle a su partager ses connaissances, ses compétences et son savoir-faire dans la valorisation des données et m'a offert un éclairage précieux sur ce métier. Enfin, je remercie l'ensemble du personnel du SGAR, pour leur accueil chaleureux.

Résumé

Ce rapport présente un projet visant à identifier en avance de phase les entreprises innovantes dans la région PACA, dans le cadre du programme France 2030. La démarche repose sur l'utilisation de l'apprentissage automatique (ML) pour détecter les entreprises ayant un fort potentiel d'innovation en combinant de plusieurs sources de données. L'étude commence par le nettoyage et l'harmonisation des données. Une analyse descriptive préliminaire a été réalisé pour comprendre les différentes variables sélectionnées pour le modèle. Ensuite, plusieurs configurations de données ont été testées avec l'algorithme SMOTENC pour équilibrer les classes. Plusieurs modèles de machine learning ont été comparés, incluant la régression logistique, les forêts aléatoires, Gradient Boosting et un modèle d'ensemble combinant ces trois algorihtmes. Les résultats montrent que le modèle d'ensemble, combinant plusieurs algorithmes, s'est révélé le plus performant pour la détection des entreprises potentielles. Des courbes d'apprentissage et des rapports de classification ont été analysés pour confirmer ces résultats.

Tables des matières

Liste des tableaux

Table des figures

Glossaire

CPER : Contrat de Plan État-Région

CRTE : Contrat de Relance et de Transition Écologique

DILA : Direction de l'Information Légale et Administrative

INSEE : Institut National de la Statistique et des Études Économiques

OCDE : Organisation de Coopération et de Développement Économiques

PACA : Provence-Alpes-Côte d'Azur

PRGR : Plan Régional de Gestion des Risques

ML : Machine Learning

SARL : Société À Responsabilité Limitée

SGAR : Secrétariat Général pour les Affaires Régionales

SIREN : Système d'Identification du Répertoire des Entreprises

SMOTE : Synthetic Minority Over-sampling Technique

SMOTENC : Synthetic Minority Over-sampling Technique for Nominal and Continuous

1 Introduction

L'innovation est un moteur essentiel du développement économique et social. Schumpeter (1935), définit l'innovation comme l'introduction réussie sur le marché d'un produit nouveau, d'un nouveau processus de fabrication ou encore d'une nouvelle forme organisationnelle de l'entreprise. Dans un monde en constante évolution, les entreprises innovantes jouent un rôle crucial en stimulant la croissance, en créant des emplois et en répondant aux défis sociétaux par des solutions novatrices. Ces entreprises, par leur capacité à innover, sont souvent à l'avant-garde de la transformation industrielle et technologique, contribuant ainsi à la compétitivité nationale et internationale.

Pour soutenir ce dynamisme, les gouvernements ont mis en place des politiques pour encourager et financer l'innovation. Ces initiatives aident les entreprises à surmonter les obstacles financiers, techniques et réglementaires, et créent un environnement favorable à l'innovation. En 2021, le président Emmanuel Macron, dévoile le plan d'investissement France 2030, qui s'inscrit dans la lignée du plan France Relance. Ce plan de 54 milliards d'euros a pour objectif de rattraper le retard de la France dans des secteurs clés, tout en créant de nouvelles filières industrielles et technologiques. Il se concentre sur l'innovation, la décarbonation, et fixe 10 objectifs prioritaires à atteindre d'ici 2030, couvrant des domaines tels que l'énergie, les transports, l'agriculture, la santé et l'exploration spatiale. La réussite de ce fonds repose en grande partie sur la capacité à détecter et à soutenir en avance de phase les entreprises innovantes susceptibles de transformer ces secteurs.

Afin d'augmenter l'impact de ce fond en région, le SGAR souhaite identifier en avance de phase les entreprises n'ayant pas déposé un dossier de soutien France 2030 mais ayant un potentiel d'innovation pouvant les rendre éligibles à un soutien financier. Compte tenu de la complexité et de la diversité de l'innovation, les méthodes traditionnelles ne sont pas toujours efficaces pour identifier ces entreprises pépites. C'est là que l'apprentissage automatique, une branche de l'intelligence artificielle, peut apporter une solution puissante. L'apprentissage automatique pourrait apporter une aide à la détection de ces entreprises et permettre à l'administration d'avoir une démarche d'accompagnement proactif vers ces entreprises. Cette approche automatisée et intelligente améliore non seulement la détection mais aussi réduit significativement le temps et les ressources nécessaires pour effectuer ces analyses, offrant ainsi un avantage stratégique considérable.

Ce rapport se concentre sur le développement d'un algorithme permettant de mettre en place cette démarche pro-active en région PACA, dans le but de maximiser l'impact du Fonds France

2030 sur la région. En utilisant des techniques avancées d'analyse de données, je développerai un modèle capable d'identifier les entreprises ayant un fort potentiel d'innovation. Ce modèle permettra de compléter les méthodes traditionnelles tels que les actions de communication auprès des partenaires institutionnelles, d'optimiser l'allocation des ressources du Fonds France 2030 et à renforcer l'écosystème d'innovation régional.

La structure de ce rapport se présente comme suit : dans un premier temps, je présenterai le contexte institutionnel de l'étude et le poste de Data Analyst, ainsi que les missions et tâches quotidiennes. Ensuite, j'aborderai la problématique spécifique de l'étude, suivie d'une revue de la littérature existence sur la détection d'entreprises innovantes et les applications de l'apprentissage automatique dans ce domaine. Enfin les résultats obtenus seront exposés, suivis de discussions et de recommandations basées sur ces résultats.

2 Contexte Institutionnel de l'étude

2.1 Présentation et structure du SGAR

Le Secrétariat Général pour les Affaires Régionales (SGAR) est un organe central dans la mise en œuvre des politiques publiques et l'organisation des services de l'Etat en région Provence-Alpes-Côte d'Azur. Il joue un rôle crucial dans la coordination des actions des différents services et des décisions gouvernementales à l'échelle régionale. En Provence-Alpes-Côte d'Azur, le SGAR travaille en étroite collaboration avec la préfecture, les préfectures départementales, les directions régionales de l'Etat (DREAL, DREETS, DRAFF, etc.), les établissements publics (ADEME, ANCT, ANRU, etc.), ainsi que divers partenaires socio-économiques et collectivités territoriales.

Le SGAR dépend du préfet de région, sous la responsabilité d'un secrétaire général M. Didier MAMIS. Il est épaulé par deux secrétaires adjoints, M. Olivier TEISSIER, en charges des politiques publiques et M. Slimane CHERIF, en charge de modernisation et des moyens. Les missions du SGAR sont multiples et couvrent un large éventail de domaines, incluant le soutien à l'économie et à l'innovation, l'aménagement du territoire, l'énergie et l'écologie, ainsi que les politiques sociales, éducatives et culturelles. Comme missions on peut citer :

- Le SGAR soutient le préfet de région dans la gestion des politiques publiques de l'État au niveau régional. Il compile les contributions des services pour préparer les décisions du préfet, supervise la rédaction des contrats et conventions avec les collectivités, et assure le suivi des actes administratifs. Au quotidien, le SGAR accompagne ou représente régulièrement le préfet lors de ses interventions publiques ;
- Le SGAR coordonne l'élaboration et la mise en œuvre du contrat de plan avec le Conseil Régional (CPER), qui constitue une feuille de route pluriannuelle pour les investissements de l'État dans la région. En lien avec les préfectures départementales, il suit également la mise en place des contrats de relance et de transition écologique (CRTE) avec les EPCI, veillant ainsi à une répartition équilibrée des ressources entre les territoires ;
- Le SGAR prépare l'attribution de nombreuses aides financières sous la responsabilité du préfet de région. Outre les financements thématiques inclus dans les contrats avec les collectivités, notamment le contrat de plan État-Région, il planifie les dotations annuelles pour soutenir l'investissement des collectivités territoriales (DSIL/DSID), le Fonds national d'aménagement et de développement du territoire (FNADT), ainsi que le « Fonds vert » destiné à accélérer la transition écologique des territoires.

- Le SGAR promeut l'innovation au sein des services de l'État en région. Grâce à son laboratoire territorial d'innovation, il teste et diffuse les bonnes pratiques issues des services ou des collectivités, et fédère les initiatives. Il fournit également les ressources nécessaires pour définir et mettre en œuvre des projets de modernisation.
- Dans le cadre des réformes de l'organisation territoriale de l'État, le SGAR constitue un appui technique crucial. Disposant d'une large expertise en ressources humaines, achats et immobilier, il aide à renforcer les capacités des services, à assurer la cohérence de leur fonctionnement, et à mener à bien leurs projets.

Les services du SGAR sont structurés en deux principaux pôles : l'un dédié au pilotage des politiques publiques, l'autre à la modernisation des services. Ces deux pôles sont soutenus par la Plateforme de Gouvernance Régionale (PFGR), qui assure la liaison, la coordination et l'organisation globale du service. Le premier pôle regroupe les chargés de mission du SGAR, qui apportent une expertise thématique et réalisent des analyses de politiques publiques pour le préfet. Ce pôle se divise en quatre sous-pôles distincts à savoir :

- ❖ La Direction Régionale aux Droits des Femmes et à l'Egalité (DRDFE)
- ❖ Cohésion Sociale Economie Emploi
- ❖ Développement Durale
- ❖ Cohésion Territoriale (où j'ai fait mon alternance)

Le second pôle est structuré autour de plateformes régionales, qui gèrent la distribution des ressources opérationnelles entre les services et encouragent la modernisation de l'administration. Ce pôle se divise également en quatre sous-sections distinctes à savoir :

- ❖ Plateforme Régionale Interministérielle d'Appui aux Ressources Humaines
- ❖ Plateforme Régionale des Achats
- ❖ Plateforme Régionales du Pilotage Budgétaire et de la Stratégie Immobilière
- ❖ Mission Modernisation Innovation

Le SGAR joue un rôle central dans la gestion des politiques publiques et la structuration des services de l'État au niveau régional. Voici comment l'organigramme du SGAR se présente :


Secrétariat particulier
Isabelle HAON
Ludivine RIBAS
Sarah BOUJAMAA
Secrétaire Général pour les Affaires Régionales de la région PACA
Didier MAMIS
PÔLE MOYENS MODERNISATION
Slimane CHERIEF
SGAR Adjoint
PÔLE POLITIQUES PUBLIQUES
Olivier TEISSIER
SGAR Adjoint
Plateforme Régionale Interministérielle d'Appui aux Ressources Humaines
Directrice
Corinne BACLET
Directrice ajointe
Estelle TAPPERO
Cheffe de projet attractivité et mobilité
Delphine GILLI
Cheffe de projet attractivité et mobilité
Agathe FERRAN
Conseillère action sociale et environnement professionnel
Elisabeth ROCHET
Conseillère formation
Karine RUGANI
Conseiller organisation du travail
Patrick VIANES
Conseillère GPRH et accompagnement des transformations
Charline HATO - MEDARIAN
Responsable de la gestion administrative et financière de l'action sociale interministérielle régionale
Laura LAVANDIER
Assistante de projets
Clémence GUERET
Plateforme Régionale des Achats
Directrice
Delphine GOBERT
Directeur adjoint
Lois PUJOLLE
Acheteuse experte fournitures courantes et services
Christine CONSOLARO
Acheteuse experte fournitures courantes et services
Nedjma ABDELLALI
Acheteuse experte fournitures courantes et services
Poste vacant
Acheteur expert immobilier
Mathieu MAYENCE
Assistant acheteur
Emma LEPORATI
Plateforme Régionale du Pilotage Budgétaire et de la Stratégie Immobilière
Directeur
Luc CLAVIER
Directrice adjointe pilotage budgétaire
Patricia GULBASDIAN
Gestionnaire BOP 354 T2
Virginie SINTES
Gestionnaire BOP 354 hors T2
Sofiane MERAH
Gestionnaire BOP SGAR politiques publiques
Vanessa PERLES
Directeur adjoint pilotage budgétaire
Pierre WERY
Gestionnaire BOP immobiliers
Nathalie MAMON
Cadre d'appui DSIL / DSID
Laurence JAUMON
Cadre d'appui chargé(e) du dispositif SPE et de la transition énergétique du parc immobilier de l'État
Edmée RINUY
Gestionnaire budgétaire en charge de la performance régionale
Poste vacant
Mission Modernisation Innovation
Cheffe de projet innovation et transformation publique
Virginie NATALE
Designer de service de la mission innovation modernisation
Poste vacant
Apprenti communication
Poste vacant
Plateforme de la Gouvernance Régionale
Directeur - Référent Mobilité / Écoresponsabilité
Philippe TRICOIRE
Directrice Adjointe
Amélie SIRVAIN
Assistante de projets
Juliette FEZAY
COORDINATION / ORGANISATION
Réunions de gouvernance - RAA
Magdalena Isabelle GIMIE
Préparation des dossiers du Préfet et du SGAR
Chloé PAUL
RH de proximité et gestion du budget de fonctionnement
Héléne WANDEL
SOUTIEN ADMINISTRATIF
Françoise AUGIER
Brigitte HAYE
Agnès BEAUDET-ROSSI
SOUTIEN DE PROXIMITE
Marc PATRICELLI
Vincent DOMENECH
Direction Régionale aux Droits des Femmes et à l'Égalité (DRDFE)
Directrice Régionale
Hélène CARON
Directrice Régionale déléguée
poste vacant
Déléguée départementale
Agnès LONCHAMP
Cadre de gestion
Poste vacant
Chargée de mission
Francisca CAMPERO
Sous - Pôle Cohésion Sociale Économie Emploi
CHARGÉS DE MISSION
Économie Emploi et Compétences
Marc GIBAUD
Référent France 2030 et Réindustrialisation
Randy KOTTI
Politique de la ville Culture Santé et Renouvellement urbain
Caroline MONNIER
Immigration Asile Jeunesse et Sport et Programmes Européens
Pauline BREMOND
Déléguée à l'information stratégique et à la sécurité économique
Claire DE GUISA
ITER
Hélène PHILIP
CHARGÉ DE PROJETS
Poste vacant
Sous - Pôle Développement Durable
CHARGÉS DE MISSION
Transports Mobilités
Magali MOINIER
Environnement Énergie et Mer
David JUNG
Agriculture Alimentation et Forêt
Jérôme HORS
CHARGÉE DE PROJETS
Rose LABEILLE
Sous - Pôle Cohésion Territoriale
CHARGÉS DE MISSION
Aménagement Logement
Cédric BASTIERI
Politiques contractuelles Fond vert
Bruno CHABAL
Assistant de projets
Yannick DJOGUERRA NARE
Massif Ruralité Tiers Lieux et Politiques Prioritaires Gouvernementales
Tessa FRECHIER MEY
Numérique
Fenitra DUPONT-RAZANAJATOVO
Apprenti numérique
Ulrich Segodo
CHARGÉE DE PROJETS
Marion DI MEO
DOTATIONS DE L'ÉTAT - FINANCES
Appui aux pôles politiques publiques et modernisation
Krystel POTHIN - Isma HANNIFER - Nolwen LEKHIAR
Secrétaire Générale de la Conférence des parties (COP) planification écologique
Muriel JOER LE CORRE
Collaborateur
Willka TINTAYA - MAMANI
Section Régionale Interministérielle de l'Action Sociale (SRIAS)
Présidente
Sylvie PUSTEL
Vice Présidente
Maryvonne GUIGONNET
Comité Consultatif Règlement Amiable (CCRA)
Présidente
Dominique BONMATI
Collaborateurs
Catherine PIETRI
Vincent DOMENECH
Commissaire à l'aménagement au développement et à la protection du Massif des Alpes
Philippe MATHERON
Commissaire à la lutte contre la pauvreté auprès du Préfet de la Région PACA
Yolande ESKENAZI
Agence Nationale de Lutte contre l'Illetrisme (ANCLI)
Corinne MATTEACCIOLI
Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI)
Délégués Régionaux
Kevin HEYDON
Célia NOWAK
Ministère des Armées
Délégué régional du ministère des Armées pour les régions Provence-Alpes-Côte d'Azur et Corse
Service de l'aménagement des territoires et de l'immobilier
Matthieu LACAILLE

2.2 Présentation du poste de Data Analyst

2.2.1 Principales responsabilité et tâches quotidiennes

En tant que Data analyst au sein du Secrétariat Général pour les Affaires Régionales (SGAR), mes responsabilités sont diversifiées et couvrent plusieurs aspects du traitement et de l'analyse des données régionales. Sous la supervision de la chargée de mission Numérique, je travaille principalement sur la collecte, la structuration, le nettoyage et l'enrichissement de données (données financières de subvention, et données sur les établissements de l'État en région), la représentation visuelle de ces données pour favoriser le pilotage et l'exploitation, la réalisation d'un algorithme d'aide à la détection des entreprises innovantes pour le programme France 2030. Dans le cadre des données financières, je récupère des données hétérogènes d'attribution de financement au niveau de chaque département ou au niveau de chaque financeur de la région, que je traite via Python sur Visual Studio Code.

Une fois les données harmonisées, je les enrichissais afin qu'au final chaque projet financé puisse être localisé au niveau d'un département, d'une intercommunalité et si possible d'une commune. J'associe également chaque projet à une thématique de politique publique et des contrats supports (CPER, CRTE). Les données sont ensuite stockées et chargées dans un outil de visualisation nommé Tableau. Cet outil, semblable à Power Bi, me permet de créer des tableaux de bords interactifs pour synthétiser et valoriser les données (annexe 1). Ces tableaux de bord sont mis à la disposition des services et des directions départementales, afin de répondre à leurs besoins de synthèse, d'analyse et de prise de décision concernant l'attribution des subventions. Des visualisations spécifiques ont été mises en place en collaboration avec le chargé de mission Politiques Contractuelles pour le pilotage du Fond Vert (fond de financement de projets en faveur de la transition énergétique) pour la préparation des comités de programmation régionaux, présidés par le préfet de Région. Mon rôle consiste à fournir des analyses précises et des visualisations claires qui soutiennent la gestion et l'allocation des ressources du Fond Vert.

Mes missions s'étendent également à la Plateforme Régionale des Ressources Humaines (PFRH), où je contribue à des projets de cartographie des services publics en PACA. Cette cartographie permet d'avoir une vision claire et structurée des différents services publics. Pour la réaliser, j'ai utilisé l'API de la DILA pour obtenir la liste complète des services publics régionaux. Après avoir nettoyé les données, elles ont été chargées dans un tableau pour créer la cartographie avec des vues à l'échelle régionale, départementale et communale (annexe 2). Enfin, je travaille étroitement avec le chargé de mission Référent France 2030 sur un projet clé : la détection des

entreprises innovantes par le biais de l'apprentissage automatique. Ce projet, qui constitue le sujet principal de mon rapport d'alternance, implique l'utilisation de techniques avancées d'analyse de données pour identifier les entreprises ayant un fort potentiel d'innovation, afin de maximiser l'impact du Fonds France 2030 en région PACA.

2.2.2 Compétences techniques et analytiques nécessaires

La réussite de ce poste nécessite plusieurs compétences techniques et analytiques. La maîtrise de Python et Spark est cruciale pour le nettoyage et le traitement des données. L'utilisation de l'environnement de développement Visual Studio Code est aussi primordial pour écrire et exécuter les scripts Python. Une bonne compréhension des techniques de manipulation des données, telles que la gestion des données manquantes, la normalisation et la transformation des données, est également indispensable. L'utilisation de Tableau pour la visualisation des données nécessite une connaissance approfondie de cet outil, notamment en ce qui concerne la création de tableaux de bord interactifs et des graphiques dynamiques, Il est également nécessaire de comprendre la donnée manipulée pour proposer des indicateurs pertinents dans les tableaux de bord. De plus, des compétences en Intelligence Artificielle notamment en apprentissage automatique sont nécessaires pour le projet de détection des entreprises innovantes. Cela inclut la compréhension des algorithmes d'apprentissage supervisé et non supervisé, la capacité à sélectionner les bonnes fonctionnalités et l'évaluation des modèles de prédiction.

2.2.3 Intégration dans les objectifs du SGAR

Mon poste de Data analyst s'intègre parfaitement dans les objectifs plus larges du SGAR. A partir des analyses de données et des visualisations que je fournis, je contribue à une meilleure vision des financements attribués sur la région, que ce soit au niveau régional, départemental et communal, et à garantir l'équité territoriale. Les tableaux de bords que je crée permettent aux services et directions départementales de suivre l'évolution des financements, d'identifier les tendances et d'optimiser l'allocation des ressources. Mon implication dans le projet de détection des entreprises innovantes par l'apprentissage automatique est directement liée aux objectifs du Fonds France 2030. En identifiant en avance de phase les entreprises à fort potentiel d'innovation, à maximiser les chances pour des entreprises du territoire de bénéficier d'aides financières en leur proposant un accompagnement au plus tôt par les services de la DREETS, et ainsi soutenir la croissance économique régionale. En travaillant sur divers projets, allant de l'analyse des financements, à la cartographie des services publics, jusqu'au soutien à une

administration pro-active, j'ai participé à la réalisation des missions de coordination et de pilotage des politiques publiques régionales du SGAR. Mon rôle contribue à assurer une gestion efficace et innovante des affaires régionales, répondant ainsi aux divers défis auxquels la région Provence-Alpes-Côte d'Azur est confrontée.

3 Problématique de l'étude : Focus sur le projet d'administration pro-active

L'innovation est un levier fondamental pour le développement économique et l'amélioration de la compétitivité régionale. Les entreprises innovantes stimulent la croissance, en créant des emplois et en apportant des solutions nouvelles aux défis sociaux, environnementaux et économiques. Au niveau national, la région PACA se distingue en occupant la 6ème place au classement des lauréats du Fonds France 2030. Dans la région, on recense 238 lauréats (identifiés par leur SIREN) sur 360 projets déposés. Cependant, l'analyse des données de la base Sirene de l'INSEE révèle la présence de 1 372 465 entreprises dans la région. Cette situation souligne la nécessité de développer une méthode innovante permettant de détecter efficacement les entreprises à fort potentiel innovant, afin de mieux les accompagner.

L'un des principaux défis est d'identifier les entreprises ayant un potentiel d'innovation élevé. Les méthodes traditionnelles, basées sur des indicateurs tels que les brevets déposés, les investissements en R&D ou les collaborations avec des institutions de recherche, peuvent être limitées et ne pas capturer toute la diversité de l'innovation. De plus, les petites et moyennes entreprises (PME) innovantes, qui, bien que dynamiques, restent souvent invisibles aux yeux de ces méthodes traditionnelles. L'utilisation de l'apprentissage automatique (ML) offre une approche potentiellement plus efficace et précise pour la détection des entreprises innovantes. Les algorithmes de ML peuvent analyser de vastes ensembles de données et identifier des modèles complexes que les méthodes traditionnelles pourraient manquer. Toutefois, la mise en œuvre de cette technologie soulève plusieurs questions :

- Quels critères et données doivent être utilisés pour former les modèles de ML afin de détecter les entreprises innovantes ?
- Comment garantit la qualité et la pertinence des données collectées et traitées ?
- Quels algorithmes de ML sont les plus appropriés pour cette tâche spécifique ?

- Comment interpréter et valider les résultats des modèles de ML pour assurer leur fiabilité et leur utilité dans le contexte de prise de décision publique ?

Une autre dimension de la problématique concerne l'alignement des résultats de l'analyse ML avec les objectifs du Fonds France 2030. Le Fonds vise à soutenir des secteurs clés tels que la transition écologique, la santé, le numérique, et l'intelligence artificielle. Par conséquent, il est nécessaire que le modèle de détection des entreprises innovantes soit capable d'identifier non seulement les entreprises innovantes en général, mais aussi celles qui sont alignées avec les priorités stratégiques du Fonds. La problématique centrale de ce projet repose sur la détection en avance de phase des entreprises susceptibles de bénéficier d'un financement dans le cadre de France 2030, afin de leur proposer un accompagnement au plus tôt dans leur demande de subvention.

4 Revue de littérature

Des études empiriques montrent que les entreprises innovantes ont un impact significatif sur la croissance économique et la compétitivité régionale. Par exemple, une étude réalisée par Adrestsch et Keilbach (2004) a démontré que les régions avec une forte densité d'entreprises innovantes avaient des taux de croissance économique plus élevés. Cette étude a analysé des données provenant de plusieurs régions européennes et a constaté que l'innovation était fortement corrélée à la création d'emplois et à l'augmentation du PIB régional. De plus, une étude faite par Czarnitzki et Delanote (2013) a montré que les PME innovantes ont une probabilité plus élevée de croissance rapide par rapport aux entreprises non innovantes. L'étude, qui a utilisé des données de l'Enquête communautaire sur l'innovation (CIS), a révélé que les entreprises qui investissent dans la R&D et introduisent des innovations de produits ou de procédés ont de meilleures performances économiques.

L'efficacité de l'apprentissage automatique pour identifier les entreprises innovantes a été démontré par plusieurs études empiriques. C'est le cas de l'étude menée par Nguyen et al. (2020), qui a utilisé des techniques d'apprentissage automatique pour analyser des données financières et non financières afin de prédire l'innovation dans les PME. L'étude a montré que les modèles d'apprentissage automatique, en particulier les réseaux de neurones et les forêts aléatoires, ont obtenu de meilleures performances par rapport aux méthodes traditionnelles. Une autre étude par Yu et al. (2019) a appliqué des techniques de clustering pour identifier des groupes d'entreprises innovantes dans le secteur technologique en chine. Les résultats ont montré que le clustering pouvait révéler des entreprises innovantes qui n'étaient pas identifiées

par des méthodes de classement traditionnelles basées sur les brevets ou les investissements en R&D.

En effet, les algorithmes de ML peuvent analyser des ensembles de données vastes et complexes plus rapidement et avec une plus grande précision que les méthodes traditionnelles. Zhang et al. (2018) ont démontré que l'utilisation de modèles d'apprentissage automatique pour prédire l'innovation des entreprises à partir de données textuelles et financières pouvait réduire le taux d'erreur de classification de 15% par rapport aux méthodes classiques. De plus, l'apprentissage automatique peut révéler des caractéristiques et des modèles d'innovation non évident, permettant ainsi une évaluation plus complète des entreprises. L'étude de Li et al. (2017) a utilisé le traitement du langage naturel (NLP) pour analyser les publications d'entreprises et identifier des indicateurs d'innovation non capturés par les données. Enfin, ces modèles d'apprentissage automatique peuvent être continuellement mis à jour et améliorer à mesure que de nouvelles données deviennent disponibles, assurant une pertinence continue. Chen et al. (2018) ont montré que les modèles de l'apprentissage automatique adaptatifs pouvaient mieux suivre les tendances d'innovation dans le secteur technologique, augmentant ainsi la précision de la détection d'innovation au fil du temps.

Plusieurs études récentes montrent l'efficacité des techniques de l'apprentissage automatique, particulière les méthodes d'ensembles et l'utilisation de SMOTE pour l'amélioration de la précision. En effet, Zhang & Ma (2023), dans leur ouvrage sur les techniques d'ensemble, explorent diverses méthodes d'ensemble comme le random forest et le boosting, qui sont efficaces pour améliorer la précision des modèles de l'apprentissage automatique. Leur recherche montre comment ces techniques peuvent être utilisées pour détecter des modèles complexes dans les données et identifier des entreprises innovantes plus précisément que les méthodes traditionnelles.

Par ailleurs, en ce qui concerne l'utilisation de SMOTE pour équilibrer les données, on note d'une part, l'étude de Hafiz & Rodiah (2021) sur la prédiction de l'attrition client dans le e-commerce dans le cas des données déséquilibrés. Après avoir réglé le problème de déséquilibre des données avec la technique SMOTE, les auteurs effectuent une modélisation à l'aide de l'algorithme Random Forest avec 02 scénarios. Les auteurs concluent qu'après utilisation du SMOTE, on obtient le meilleur modèle pour obtenir la bonne valeur de précision sur les données d'entraînement et les données test. D'autres part, Wongvorachan et al. (2023) ont comparé plusieurs techniques de suréchantillonnage, y compris SMOTE, pour gérer les classifications déséquilibrées dans le domaine de l'éducation. Leur étude a démontré que SMOTE améliore

significativement la performance des modèles prédictifs, en augmentant la précision des classes minoritaires sans introduire de biais. Cette approche peut être directement appliquée à la détection des entreprises innovantes où les cas de succès peuvent être rares comparativement aux autres entreprises.

5 Présentation des données

5.1 Source des données

Les données utilisées dans cette étude pour la détection des entreprises innovantes dans la région PACA, proviennent de sources différentes. Ces données proviennent de bases de données publiques et de partenariats avec d'autres administrations. Voici une description détaillée de chaque source de données et des critères de sélection utilisés pour garantir leur pertinence et leur fiabilité.

- **Listes des projets des lauréats France 2030**

Cette base de données contient les informations sur les projets financés dans le cadre du programme France 2030. Les projets couvrent divers secteurs innovants et permettent d'identifier les entreprises bénéficiant du soutien de ce fonds. Les données comprennent des informations pour identifier l'entreprise porteur du projet, la nature et l'objectif des projets, les montants alloués. Cette source est importante car elle permet d'identifier les entreprises innovantes soutenues par le programme France 2030, offrant ainsi une vision claire des secteurs où l'innovation est la plus dynamique.

- **Données du crédit d'impôt-recherche (CIR) 2020**

Les données du CIR offrent des informations sur les entreprises qui bénéficient d'incitations fiscales pour la recherche et le développement. Ces entreprises sont souvent à la pointe de l'innovation, et l'analyse de ces données permet de repérer celles qui investissent massivement en R&D. Les critères de sélection pour cette source incluent la pertinence des entreprises par rapport à l'innovation technologique et leur capacité à bénéficier du CIR.

- **Fichier INSEE des établissements de PACA**

Le fichier INSEE fournit une base de données exhaustive des établissements situés dans région PACA. Cette source est utilisée pour obtenir des informations démographiques et économiques

de base sur les entreprises, telles que la taille, le secteur d'activité et la localisation. Elle permet aussi de connaitre les autres entreprises non lauréates de la région. Les critères de sélection comprennent la mise à jour régulière des données et la couverture complète des établissements dans la région.

- **Fichier Dealroom : lévées de fonds des entreprises**

Dealroom est une plateforme de données sur les startups et les entreprises technologiques. Le fichier dealroom contient des informations détaillées sur les startups en PACA, incluant leur financement, leur activité et leurs performances. Cette source est essentielle pour identifier les jeunes entreprises à fort potentiel innovant. Les critères de sélection pour cette source incluent la pertinence des données pour les secteurs technologiques et innovants.

- **Liste des projets refusés ou ajournés France 2030**

Cette liste comprend les projets qui ont été soumis pour financement mais n'ont pas été acceptés ou ont été ajournés. L'analyse de ces données permet de comprendre les critères de sélection du fonds France 2030 et d'identifier les entreprises qui, bien que n'ayant pas reçu de financement, sont engagées dans des activités innovantes.

- **Listes des projets déposés France 2030**

Cette source compile tous les projets soumis dans le cadre de France 2030, qu'ils aient été acceptés, refusés ou en attente. Ces données offrent une vue complète des entreprises qui cherchent à innover et à obtenir le financement France 2030. Les critères de sélection incluent la complétude des informations sur les projets.

En combinant ces différentes sources, l'étude vise à obtenir une vue d'ensemble précise et complète et l'écosystème des entreprises innovantes en PACA.

5.2 Sélection et description des variables

Cette section décrit les différentes variables clés utilisées dans notre étude pour construire notre modèle de détection des entreprises innovantes. Dans notre étude, nous avons choisi d'utiliser les numéros SIREN au lieu des numéros SIRET pour identifier les entreprises. Le SIREN est un numéro unique attribué à chaque entreprise, tandis que le SIRET est un numéro attribué à chaque établissement de l'entreprise. Le SIREN permet donc de regrouper les informations au niveau de l'entreprise entière, ce qui est plus pertinent pour notre étude. La variable dépendante ici appelée

A_deposer, prendre la valeur 1 si l'entreprise a déposé au moins un projet pour solliciter la subvention France 2030, que le projet soit accepté, refusé ou ajourné, et 0 sinon. Le résultat final de la modélisation est d'identifier ces entreprises qui n'ont jamais sollicité les subventions France 2030 mais que le modèle identifie comme appartenant à la catégorie 1.

Variables	Type	Description	Modalité
A_deposer	Binaire	Indicateur binaire de soumission de projet pour financement France 2030	0 = Non, 1 = Oui
DEP	Catégoriel	Département où est situé l'entreprise	04, 05, 06, 13, 83, 84
CJ	Catégoriel	Catégorie juridique de l'entreprise	Différentes catégories juridiques comme SARL, SAS, etc.
APET	Catégoriel	Activité principale de l'entreprise	
A_lever_fond ou A_declarer_R&D	Binaire	Indicateur binaire de l'intention de lever des fonds ou de déclarer la R&D	0 = Non, 1 = Oui
EFFEQTP_g	Quantitatif	Effectif à temps plein global	Numérique
salaire_moy	Quantitatif	Salaire moyen des employés	Numérique

_TABLE 1 - Présentation des variables

5.3 Traitement des données

Le traitement des données est une étape cruciale pour garantir la qualité et la fiabilité des résultats obtenus à partir des analyses. Voici les étapes principales du processus de traitement des données, ainsi que les techniques et outils utilisés pour chacune d'elles.

5.3.1 Nettoyage des données.

Le nettoyage des données. Est essentielle pour éliminer les erreurs, les incohérences et les doublons. Les étapes suivantes ont été suivies.

- **Suppression des doublons** : nous avons utilisé la fonction ***drop_duplicates*** de Pandas pour identifier et supprimer les enregistrements dupliqués.

- **Correction des erreurs** : Dans ce cas, nous procédons à une vérification manuelle. Et à une correction des erreurs typographiques ou des incohérences dans les noms des variables et les valeurs.
- **Gestion des valeurs manquantes :** Dans cette étude. Nous n'avons pas eu besoin de gérer les valeurs manquantes car les ensembles de données utilisés étaient complets. La qualité des données collectées était suffisante pour éviter les problèmes liés aux valeurs manquantes, ce qui a simplifié le processus de nettoyage des données.

5.3.2 Prétraitement des données.

Pour s'assurer que les variables sont à une échelle comparable, les données ont été standardisées.

- **Standardisation :** cette technique permet de centrer et de réduire les variables afin qu'elles aient une moyenne de 0 et un écart-type de 1. Elle est particulièrement utile dans le contexte des modèles de machine Learning, car elle permet de garantir que chaque variable contribue de manière équivalente à l'algorithme d'apprentissage, sans qu'une variable ayant une grande échelle n'influence disproportionnellement les résultats. En utilisant ***StandardScaler*** de Scikit-learn, chaque valeur de variable est transformée selon la formule suivante :

$$Z = \frac{(X - \mu)}{\sigma}$$

  Où X est la valeur originale, μ est la moyenne des valeurs de la variable de la variable et σ est l'écart-type des valeurs de la variable.
- **Traitement des valeurs aberrantes :** les valeurs aberrantes (ou outliers) peuvent avoir un impact significatif sur les analyses. Cependant dans le contexte de la détection d'entreprises innovantes, les valeurs aberrantes sont particulièrement importantes car elles peuvent représenter des cas d'innovation exceptionnelle ou de performance financière inhabituelle. Dans cette étude, les outliers ne seront donc pas retirés.
- **Transformation des variables** : certaines variables nécessitaient des transformations pour être utilisables dans les modèles d'apprentissage automatique. C'est le cas des variables catégoriques comme CJ, DEP et APET qui ont été transformées en variables numériques muettes à l'aide de la technique d'encodage de Pandas, ***get_dummies***. D'autres variables ont été dérivées des données existantes pour mieux représenter les caractéristiques d'intérêt. Nous avons créé les variables comme le carré des variables EFFEQTP_g et salaire_moy ainsi que l'interaction entre ces deux variables.

5.3.3 Gestion du déséquilibre des données

Dans ce projet, notre variable d'intérêt, "A_deposer", montre un déséquilibre significatif entre les deux classes. Pour gérer ce problème, nous avons employé la méthode SMOTE (Synthetic Minority Over-sampling Technique). SMOTE est une technique qui génère des exemples synthétiques pour la classe minoritaire afin de rééquilibrer les classes. Au lieu de simplement copier les instances minoritaires existantes, SMOTE crée de nouvelles observations en combinant des caractéristiques de plusieurs exemples proches dans l'espace des données. Cette approche permet de rendre la distribution des données minoritaires plus homogène, en enrichissant leur diversité, ce qui améliore la performance des modèles d'apprentissage automatique. Concrètement, SMOTE peut être décomposé en cinq étapes principales :

- Choix d'un vecteur de caractéristique de notre classe minoritaire que nous appellerons vc ;
- Sélection des k-voisins les plus proches (k=5 par défaut) et choix de l'un d'eux au hasard que l'on appellera pv ;
- Calcul de la différence pour chaque valeur caractéristique (feature value) i, vc[i] – pv[i] et multiplication de celle-ci par un nombre aléatoire entre [0, 1] ;
- Ajout du résultat précédent à la valeur de la caractéristique i du vecteur vc afin d'obtenir un nouveau point (une nouvelle donnée) dans l'espace des caractéristiques ;
- Répétition de ces opérations pour chaque point de données de la classe minoritaire ;

Au lieu d'utiliser SMOTE pour équilibrer les données, nous avons opté pour une variante appelée SMOTENC (SMOTE-Nominal Continuous), spécialement conçue pour gérer des données mixtes qui incluent à la fois des variables numériques et catégorielles. Étant donné la nature mixte de nos données, SMOTENC est bien mieux adapté à notre contexte. La principale distinction entre ces deux algorithmes réside dans la manière dont ils calculent les distances pour identifier les voisins les plus proches et génèrent des observations synthétiques, en intégrant les variables catégorielles dans SMOTENC. Avec SMOTENC, le calcul de la distance entre deux individus s'effectue en deux phases lorsqu'il s'agit de trouver les k voisins les plus proches de l'observation minoritaire initiale :

A. Calcul de la distance euclidienne à partir des variables numériques
B. Pour chaque variable catégorielle, augmentation de la distance selon les deux cas suivants :
    - Si les deux individus ont la même modalité pour cette catégorielle, alors on augmente la distance de 0

- Si les 2 individus ont une modalité différente, alors on augmente la distance de D.

D est défini comme la médiane des écarts-types des variables numériques. Cette valeur permet d'avoir une distance qui soit cohérente entre les variables numériques et catégorielles.

La formule suivante résume la distance entre deux individus avec une combinaison de variables numériques et catégorielles :

$$dist(X_i, X_j) = \sqrt{\sum_{k=1}^{K_{num}} \left(X_k^i - X_k^j\right)^2} + D \sum_{k=K_{num}+1}^{K_{cat}} 1_{\left\{X_k^i = X_k^j\right\}}$$

Où

$$D = median_{1 \le k \le K_{num}}(\sigma_k)$$

Une fois les plus proches voisins identifiés, la création de l'individu synthétique est réalisée de la façon suivante :

- Pour les variables numériques, le SMOTENC calcule les valeurs de la même façon que le SMOTE
- Pour chaque variable catégorielle, la modalité du nouvel individu correspond à la modalité la plus fréquente parmi ses k plus proches voisins.

6 Analyse exploratoire des données

6.1 Analyse univariée

L'analyse des variables quantitatives de notre base, à savoir l'effectif temps plein des entreprises (**EFFEQTP_g**) et le salaire moyen par employé (**salaire_moy**) révèle une grande disparité parmi les entreprises. En moyenne, chaque entreprise compte environ 14 employés à temps plein, mais cette moyenne est influencée par quelques très grandes entreprises, comme le montre l'écart-type élevés de 104.33 et la médiane beaucoup plus basse de 4. Les 25% des plus petites entreprises comptent moins de 2 employés, tandis que les 25% des plus grandes entreprises ont jusqu'à 8 employés ou plus. L'entreprise le plus grand compte quant à elle 8283 employés.

En ce qui concerne les salaires, le salaire moyen par employé est de 35396.35 euros, mais là encore, il y a une grande variabilité (écart-type de 21152.24). Les salaires vont d'un minimum de 4177.98 euros à un maximum de 607340.50 euros, indiquant la présence de quelques entreprises offrant des salaires très élevés. La moitié des entreprises ont des salaires moyens inférieurs à 30404.15 euros, tandis que 25% des entreprises ont des salaires moyens inférieurs à 24478.95 euros et 25% ont des salaires supérieurs à 39904.14 euros. Ces données mettent en évidence des différences marquées en termes de taille et de rémunération des entreprises. De même le graphique de distribution, en annexe 3 du document, nous montre que la distribution de ces variables ne suit pas une loi normale.

	EFFEQTP_g	salaire_moy
count	20446	20446
mean	14	35396.35
std	104.33	21152.24
min	1	4177.98
25%	2	24478.95
50%	4	30404.15
75%	8	39904.14
max	8283	607340.5

TABLE 2 - Présentation des variables quantitatives

La répartition de la variable **A_deposer** présentée par la figure 1 nous indique que près de 99% des entreprises PACA de notre base de données n'a jamais déposé un dossier auprès de France 2030. En analysant aussi la répartition des entreprises suivant la variable **A_lever_fonds ou A_declarer_R&D**, on relève aussi un déséquilibre entre les deux catégories d'entreprises. Les entreprises bénéficiant d'incitations fiscales pour la recherche et le développement ne représentent qu'environ 9% des données globales de la base de données. Ceci nous permet de visualiser déjà le déséquilibre de données auquel nous avons eu à faire face dans cette étude.

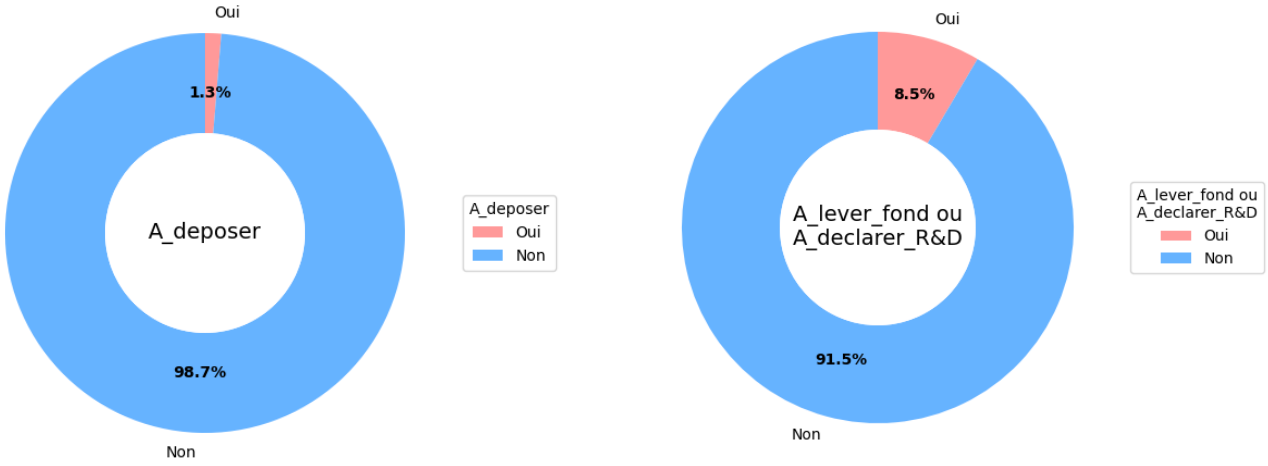


Figure 1 - Répartition des entreprises selon les variables A_deposer et A_lever_fond ou A_declarer_R&D

La figure 2 présente la répartition des entreprises par département. Nous constatons que le département du Bouches-du-Rhône (13) compte le plus grand nombre d'entreprise, suivi des départements des Alpes Maritimes (06) et le Var (83). Les départements des Hautes-Alpes (05) et des Alpes-de-Haute-Provence (04) sont ceux qui ont le faible nombre d'entreprise.

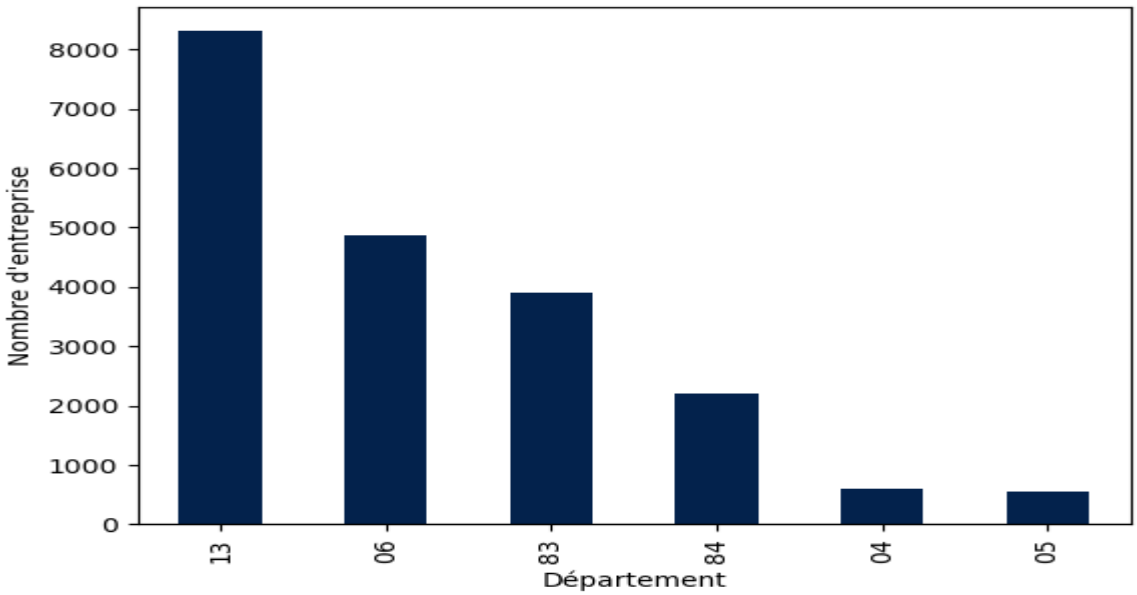


Figure 2 - Répartition des entreprises par département

La figure 3 présente le Top 10 de la répartition des APET (Activités Principale de l'Entreprise) des entreprises de la base de données. Nous visualisons ici le Top 10 car nous avons dans la base 86 catégories d'APET. Nous remarquons que les entreprises ayant comme APET la restauration traditionnelle domine très largement les autres catégories d'entreprises dans les données.

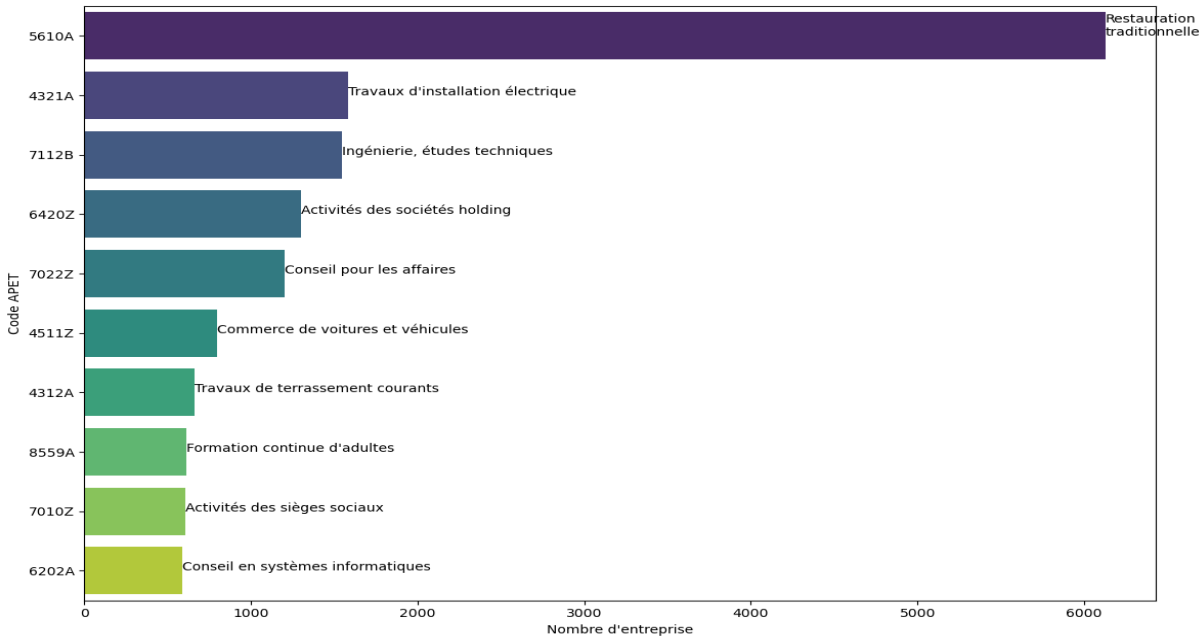


Figure 3 - Répartition des entreprises selon l'APET

La figure 4 présente le Top 5 de la répartition des entreprises par Catégorie Juridique. Dans la même logique que la figure précédente nous avons choisit d'afficher le Top 5 pour plus de clarté dans l'affichage des données. Nous constatons que les entreprises de la catégorie juridique Ste par action et S.A.R.L. dominent très largement l'ensemble des autres catégories de la base.

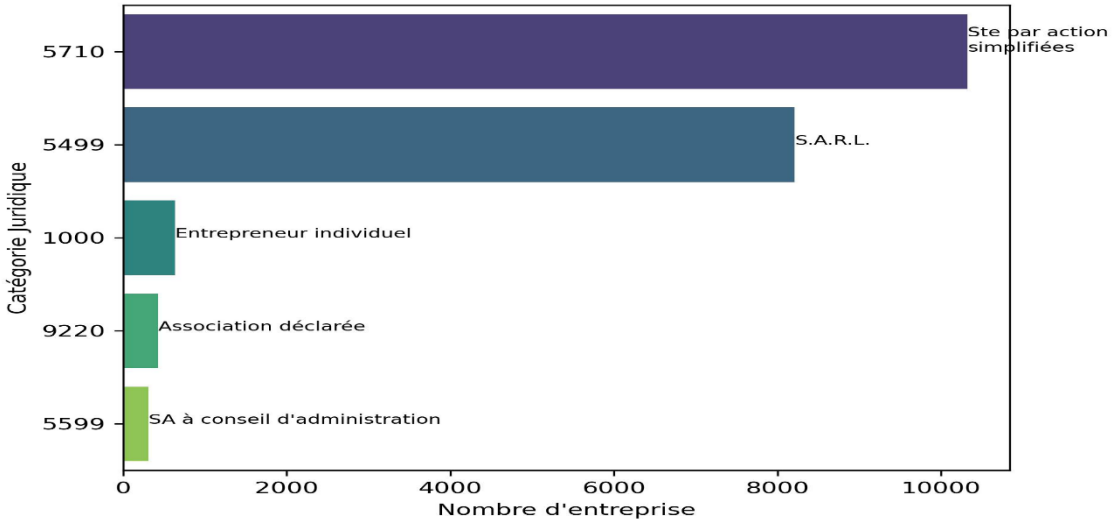


Figure 4 - Répartition des entreprises selon la catégorie juridique

6.1.1 Analyse bivariée

Nous cherchons à analyser la relation entre certains variables et la variable **A_deposer** (variable d'intérêt)

La figure 5 présente la répartition en pourcentage des entreprises ayant déposés au moins un projet à France 2030 suivant le département. Nous constatons que dans tous les départements, le part des entreprises ayant déposés à France 2030 est très faible. Ceci justifie l'importance de cette étude, qui a pour but d'augmenter le nombre d'entreprise bénéficiant d'un accompagnement à France 2030.

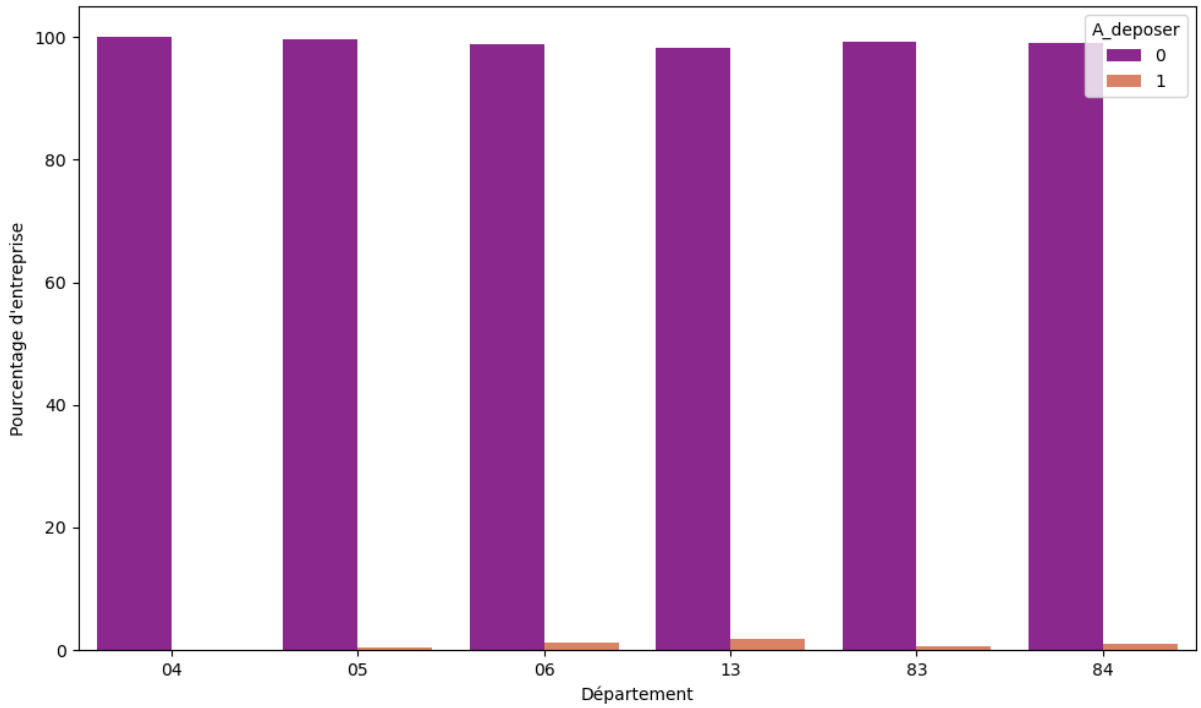


Figure 5 - Répartition des entreprises par département et par statut de dépôt

La figure 6 présente la répartition des entreprises suivant les variables **A_deposer et A_lever_fond ou A_declarer_R&D**. D'une part, on constate même si c'est dans une faible proportion, qu'il y a des entreprises qui n'ont pas levées de fonds ou déclarer des données de recherche et développement mais qui ont quand même déjà déposer à France 2030. D'autre part, la proportion d'entreprise ayant déjà déposer à France 2030, reste très faible dans les 2 catégories.

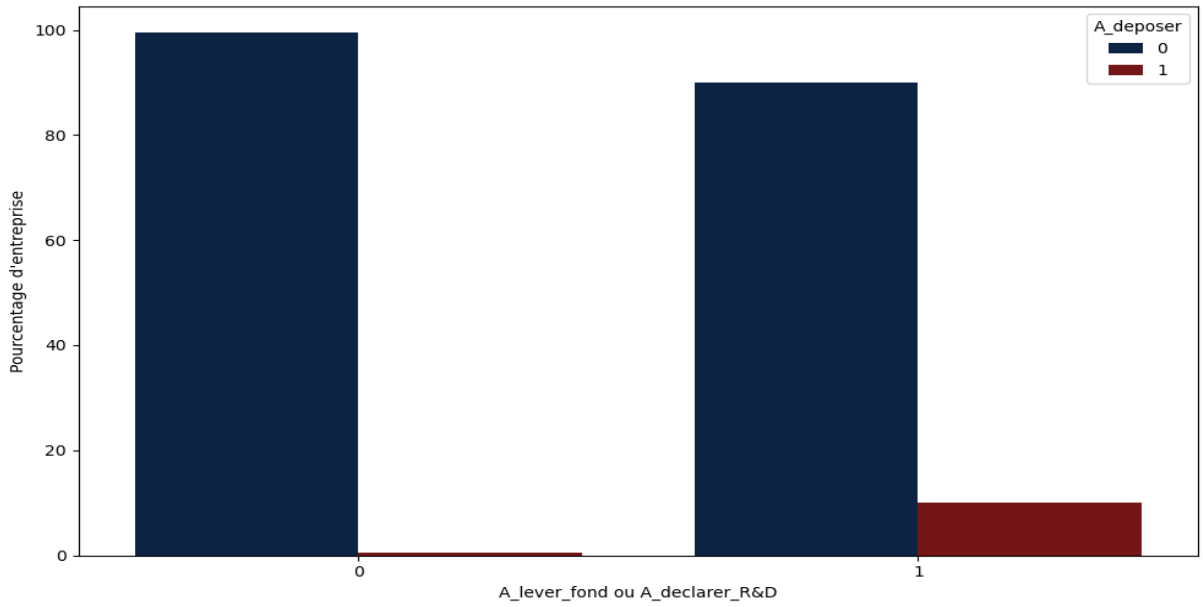


Figure 6 - Répartition des entreprises par statut de dépôt et de levée de fond ou R&D

Le graphique de corrélation basé sur la méthode de Spearman, nous montre que les variables quantitatives de notre base à savoir : l'effectif temps plein (EFFEQTP_g) et le Salaire moyen par employé (salaire_moy) ont une corrélation positive mais faible, soit de 0.34.

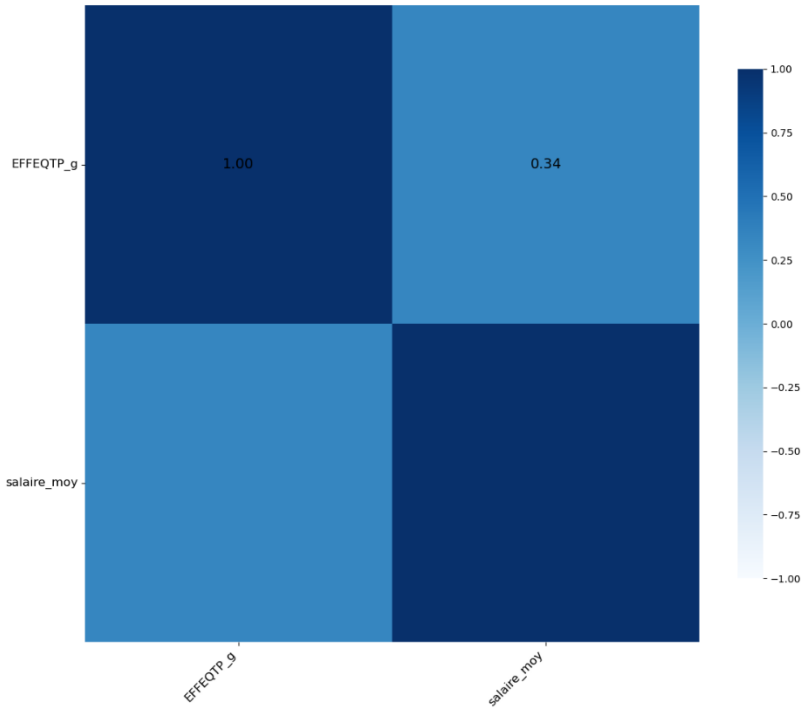


Figure 7 - Matrice de corrélation

7 Présentation des résultats

Dans cette section, nous allons présenter les résultats des analyses et des modèles d'apprentissage automatique. Nous utiliserons des tableaux, des graphiques et des visualisations pour illustrer les découvertes. Nous comparerons les performances des différents modèles que nous avions testés et expliquerons pourquoi certains modèles se sont révélés plus efficaces que d'autres pour la détection d'entreprises innovantes.

7.1 Normalisation des données

Dans cette étude, nous sommes confrontés à un défi majeur lié au déséquilibre des données. Pour tirer pleinement parti de l'algorithme de suréchantillonnage SMOTENC, il est essentiel de normaliser les données. Cela permet de réduire l'impact des valeurs extrêmes et améliorer les performances de l'algorithme.

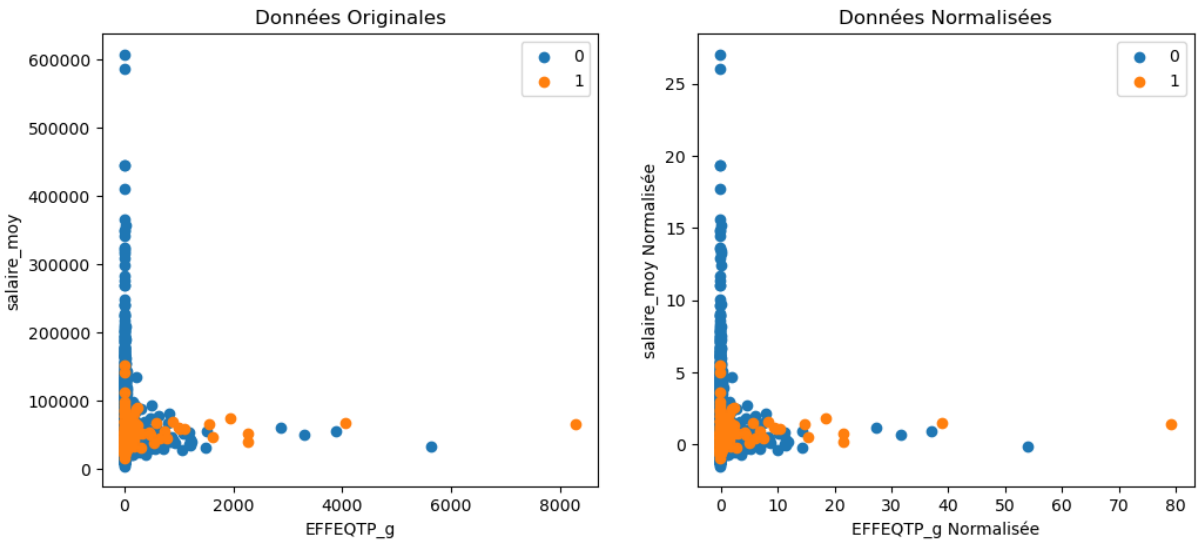


Figure 8 - Comparaison des données originales et des données normalisées

Dans le graphique de gauche, nous observons une forte concentration de points près de l'origine (0,0) avec quelques points dispersés sur les axes *'EFFEQTP_g'* et *'salaire_moy'*. Cette concentration indique que la plupart des entreprises ont un effectif et un salaire moyen relativement bas, avec quelques valeurs extrêmes. Les valeurs extrêmes peuvent influencer les performances des algorithmes d'apprentissage automatique, en particulier ceux qui utilisent des mesures de distance comme le k-NN (utilisé dans SMOTE). Le graphique de droite montre les mêmes données après normalisation. Cette transformation permet de réduire l'impact des valeurs extrêmes, de faciliter la convergence des algorithmes d'apprentissage automatique,

d'améliorer la comparabilité des variables qui étaient initialement sur des échelles différentes. Après cette normalisation, nous constatons que les données sont plus centrées autour de l'origine, avec des valeurs plus uniformément réparties sur les axes. Les variations extrêmes sont atténuées, ce qui nous permet ensuite d'appliquer l'algorithme d'équilibrage des données.

7.2 Application de SMOTENC

Pour bien gérer le déséquilibre des données et contrôler le nombre d'observations synthétiques générées par SMOTENC afin de garantir leur qualité, nous avons testé quatre configurations de données résumées dans le tableau ci-dessus. Dans deux de ces configurations nous avons appliqué des techniques de sous échantillonnage pour réduire la classe majoritaire. Pour s'assurer que cette classe majoritaire réduite contienne toujours suffisamment d'information pour l'apprentissage des modèles, nous avons utilisé la méthode de Centroids de Clusters pour sélectionner des représentants significatifs de la classe majoritaire.

	Classe 0	Classe 1	Total
Jeu de données d'origine	20180	266	20446
(équivalent à **Pourcentage SMOTENC = 0**)	0.99%	0.01%	
Pourcentage SMOTENC = 0	500	266	766
Cluster Centroids = 500	0.66%	0.34%	
Pourcentage SMOTENC = 34%	7000	7000	14000
Cluster Centroids = 7000	50%	50%	
Pourcentage SMOTENC = 50%	20180	10090	30270
	66%	34%	
Pourcentage SMOTENC = 100%	20180	20180	40360
	50%	50%	

TABLE 3 - Synthèse Application SMOTENC

Le jeu de données initial présente un fort déséquilibre entre les classes, avec une majorité écrasante pour la classe 0. Dans la première configuration, en appliquant la méthode des

centroïdes de Clusters, nous avons réduit la classe majoritaire à 500 observations, augmentant ainsi la proportion de la classe minoritaire. La deuxième configuration est un sous échantillonnage plus modéré de la classe majoritaire et une augmentation de la classe minoritaire avec SMOTENC. Nous avons atteint dans ce cas un équilibre parfait entre les classes. Ensuite, dans la troisième configuration, nous avons conservé toutes les observations de la classe majoritaire et augmenté la classe minoritaire à 34% du total avec SMOTENC. Enfin nous avons doublé la classe minoritaire comparé au précédent, pour obtenir un équilibre parfait entre les classes. Ces configurations montrent différentes approches pour gérer le déséquilibre des données. L'annexe 4 montre une représentation graphique des données dans chaque cas par rapport aux données originales. Dans chaque configuration de données, nous avons entrainé trois modèles de machine learning à savoir : La régression logistique, les forêts aléatoires, le Boosting de Gradient. Nous avons également créé un modèle d'ensemble combinant ces trois modèles. L'objectif est de comparer les performances des modèles dans chaque configuration de données et d'identifier celle qui offre les meilleures prédictions tout en minimisant les effets de surajustement.

7.3 Présentation des modèles

- Régression Logistique (Logit Binaire)

Le logit binaire est un modèle très utilisé dans les cas des variables dépendantes qualitatives et prenant deux valeurs. Dans un logit binaire, les modalités de la variable à expliquer sont 0 ou 1 et indique la réalisation ou non d'un évènement pour l'individu. La régression logistique utilise une fonction logistique pour modéliser la probabilité qu'un évènement se produise en fonction de plusieurs variables explicatives. L'équation du modèle est la suivante :

$$P(Y = 1/X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

$P(Y = 1/X)$: Probabilité que la classe soit 1

$\beta_0, \beta_1, \ldots\ldots\ldots \beta_n$: Coefficients du modèle

Bien qu'elle suppose une relation linéaire entre les variables explicatives et la log-odds, elle est rapide à entraîner et facile à comprendre. Son principal atout réside dans la clarté des coefficients qui indiquent l'importance relative de chaque variable, mais elle peut être sensible aux valeurs extrêmes

- Forêt aléatoire (Random Forest)

Une forêt aléatoire est un ensemble d'arbres où chaque arbre dépend d'une collection de variables aléatoires et est basée sur les arbres de décision. Il s'agit d'une amélioration du Bagging. La forêt aléatoire construit de nombreux arbres, chacun d'entre eux étant construit en tenant compte d'un sous ensemble aléatoire d'observations et d'un sous ensemble aléatoire de variables. Dans cet algorithme, pour une observation donnée à prédire, la prédiction d'un arbre s'obtient en faisant passer cette observation par les branches, en fonction des conditions d'écrêtage. La prédiction de la forêt aléatoire correspond à la moyenne des prédictions de chacun de ces arbres. Ce modèle offre plusieurs avantages, notamment la réduction du risque de surajustement grâce à la diversité des arbres et la capacité à gérer de grandes quantités de données et de nombreuses caractéristiques. Dans notre étude, nous décidons d'optimiser les paramètres du modèle comme pour tous les autres modèles. Nous utilisons la validation croisée par stratification pour trouver les meilleurs paramètres et ces paramètres sont utilisés pour le modèle final.

- Le boosting de Gradient (Gradient Boosting)

Le Boosting de Gradient est une technique d'ensemble qui améliore les performances des modèles en ajoutant des arbres de décision successifs, chaque nouvel arbre corrigeant les erreurs des arbres précédents. L'un des algorithmes les plus populaires pour cette méthode est XGBoost. Le modèle final est une combinaison pondérée de tous les arbres, construits de manière séquentielle pour minimiser une fonction de coût. Cette approche est particulièrement efficace pour capturer des relations complexes entre les caractéristiques et la variable cible, offrant ainsi d'excellentes performances pour une large gamme de tâches de classification et de régression. De plus, la régularisation intégrée dans les algorithmes de Boosting de Gradient aide à prévenir le surajustement. Cependant, cette méthode peut être sensible aux valeurs aberrantes si les paramètres ne sont pas correctement réglés, et elle nécessite un temps de formation plus long en raison de la nature séquentielle de l'algorithme.

7.4 Choix de la meilleure configuration des données

Pour déterminer la configuration optimale des données permettant une détection précise et fiable des entreprises innovantes, nous avons comparé plusieurs configurations en utilisant le modèle de régression logistique. Ce modèle a été choisi en raison de sa simplicité, de sa robustesse et de son interprétabilité, ce qui en fait un excellent point de départ pour comparer différentes configurations de données. En outre, les courbes d'apprentissage générées par ce modèle permettent de visualiser clairement les effets de surajustement (overfiiting) et de sous ajustement (underfitting) sur nos jeux de données. Le graphe ci-dessous montre une

comparaison des courbes d'apprentissage pour un modèle de régression logistique dans les quatre configurations.

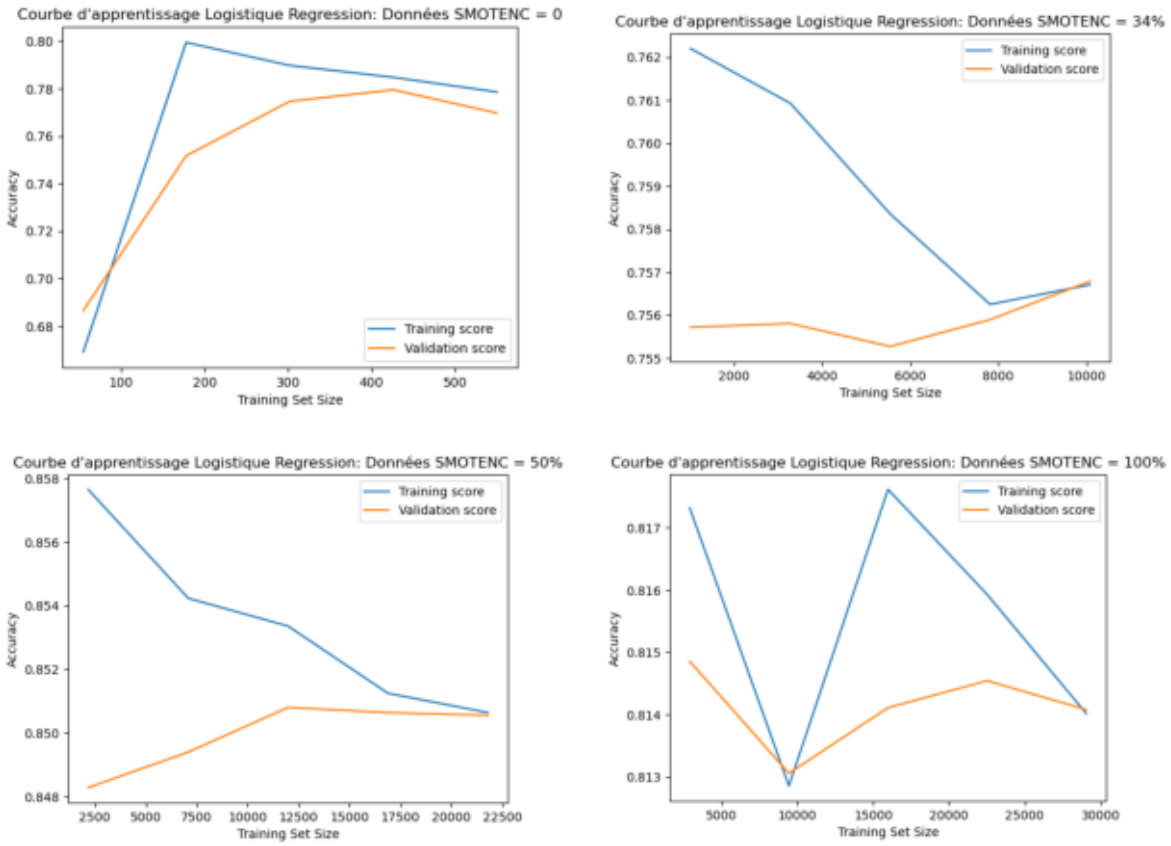


Figure 9 - Courbes d'apprentissage des modèles

De l'analyse de ce graphique, nous constatons que dans la configuration où aucune donnée synthétique n'a été ajoutée (SMOTENC =0), le score d'entrainement atteint environ 0.78 avec 200 observations, puis se stabilise autour de 0.76. Le score de validation est en revanche constamment inférieur au score d'entrainement. Les scores d'entrainement et de validation sont proches mais bas, indiquant un sous ajustement du modèle. On peut déduire que ce modèle n'apprend pas suffisamment bien des données. Dans la deuxième configuration (Données SMOTENC= 34%), le score d'entrainement commence élever (~0.76) mais décroit légèrement avec la taille de l'échantillon tandis que le score de validation reste constant (~0.75) légèrement en dessous du score d'entrainement. Cette réduction de l'écart entre les scores indique une légère amélioration, mais les scores restent relativement faibles. L'ajout des données synthétiques a aidé réduire le déséquilibre, mais le modèle n'est toujours pas optimisé. En analysant la troisième configuration (SMOTENC = 50%), on remarque que les scores sont plus élevés et proches, suggérant une bonne performance du modèle avec cette configuration. Cette

configuration semble être bien équilibrée réduisant à la fois le surajustement et le sous-ajustement. Enfin, dans la dernière configuration de données (SMOTENC = 100%), le score d'entrainement est initialement élevé (~0.817) mais montre une forte variabilité tandis que le score de validation est plus stable mais légèrement inférieur. Bien que les scores soient élevés, la forte variabilité indique une possible instabilité du modèle avec ce niveau d'équilibrage. On en déduit que l'équilibrage total avec SMOTENC semble introduire une variabilité dans les résultats, ce qui peut ne pas être idéal pour la stabilité du modèle. De ces analyses, On conclut que la configuration avec SMOTENC à 50% semble offrir un bon équilibre, avec des scores d'entrainement et de validation proches et élevés, indiquant un bon compromis entre biais et variance. Les résultats de l'estimation des modèles qui seront présentés, seront basés sur cette configuration des données.

7.5 Evaluation des modèles

Nous présentons ici les résultats de l'évaluation de plusieurs modèles de machine learning appliqués à notre jeu de données déséquilibré. Les modèles évalués comprennent la régression logistique, les forêts aléatoires, le gradient boosting et les méthodes d'ensemble. Pour chaque modèle, nous avons calculé les métriques de précision, de rappel, de F1-score, et de précision globale afin d'identifier le modèle offrant les meilleures performances pour la détection des entreprises innovantes.

Model	Class	Precision	Recall	F1-Score
Logistic Regression	0	0.86	0.96	0.89
	1	0.83	0.70	0.76
Random Forest	0	0.94	0.93	0.93
	1	0.86	0.88	0.87
Gradient Boosting	0	0.90	0.93	0.92
	1	0.86	0.80	0.82
Ensemble methods	0	0.93	0.93	0.93
	1	0.86	0.86	0.86
Accuracy	0.85 (Logistic Regression) 0.91 (Random Forest) 0.89 (Gradient Boosting) 0.91 (Ensemble methods)			

TABLE 4 - Rapport de classification

La régression logistique montre une bonne performance pour la classe 0, avec une précision et un rappel élevé. Pour la classe 1, bien que la précision soit élevée, le rappel est relativement bas, indiquant que le modèle a tendance à manquer certains échantillons de classe 1. Le F1-score de la classe 1 est inférieur à celui de la classe 0, suggérant une meilleure performance globale pour la détection de la classe majoritaire. La forêt aléatoire montre une performance équilibrée entre les deux classes, avec des scores de précision et de rappel élevés pour les deux classes. Les scores F1 indiquent que le modèle traite bien le déséquilibre des classes, détectant efficacement les classes minoritaires sans perdre de précision. La précision globale élevée (0.91) confirme la robustesse de ce modèle. Le gradient boosting montre une bonne performance pour la classe 0, similaire à la forêt aléatoire, mais avec un rappel légèrement inférieur pour la classe 1. Le F1-score pour la classe 1 est inférieur à celui de la forêt aléatoire, indiquant que le modèle peut rencontrer des difficultés à détecter tous les échantillons de la classe minoritaire. La précision globale (0.89) reste élevée, montrant que ce modèle est performant mais légèrement moins équilibré que la forêt aléatoire. Les méthodes d'ensembles combinent les forces des différents modèles pour atteindre des performances équilibrées. Les scores F1 équilibrés pour les deux classes montrent que cette approche permet de bien traiter le déséquilibre des classes. La précision globale élevée (0.91) indique une performance robuste et fiable pour la classification. En conclusion, pour une détection précise et équilibrée des entreprises innovantes, les méthodes d'ensemble et les forêts aléatoires sont meilleurs.

7.6 Comparaison de la calibration des probabilités

Dans les modèles prédictifs, et plus particulièrement dans les tâches de classification, il est crucial non seulement de faire des prédictions précises mais aussi d'estimer des probabilités fiables pour ces prédictions. La courbe de calibration des probabilités permet de visualiser si les probabilités prédites dans un modèle reflètent la véritable probabilité des évènements. Cela est particulièrement important dans notre contexte car bous cherchons à identifier les entreprises pour lesquels le modèle prédit une forte probabilité d'avoir déposé à France 2030 mais qui n'ont jamais déposé un projet. Une calibration correcte du modèle garantit que les probabilités prédites correspondent bien aux résultats réels, augmentant ainsi la fiabilité et l'utilité du modèle dans les applications pratiques. Le graphique de calibration des probabilités ci-dessous compare les probabilités prédites des quatre modèles entrainés avec les probabilités réelles. Plus la courbe d'un modèle est proche de la diagonale (parfaitement calibrée) meilleures sont les estimations de probabilités.

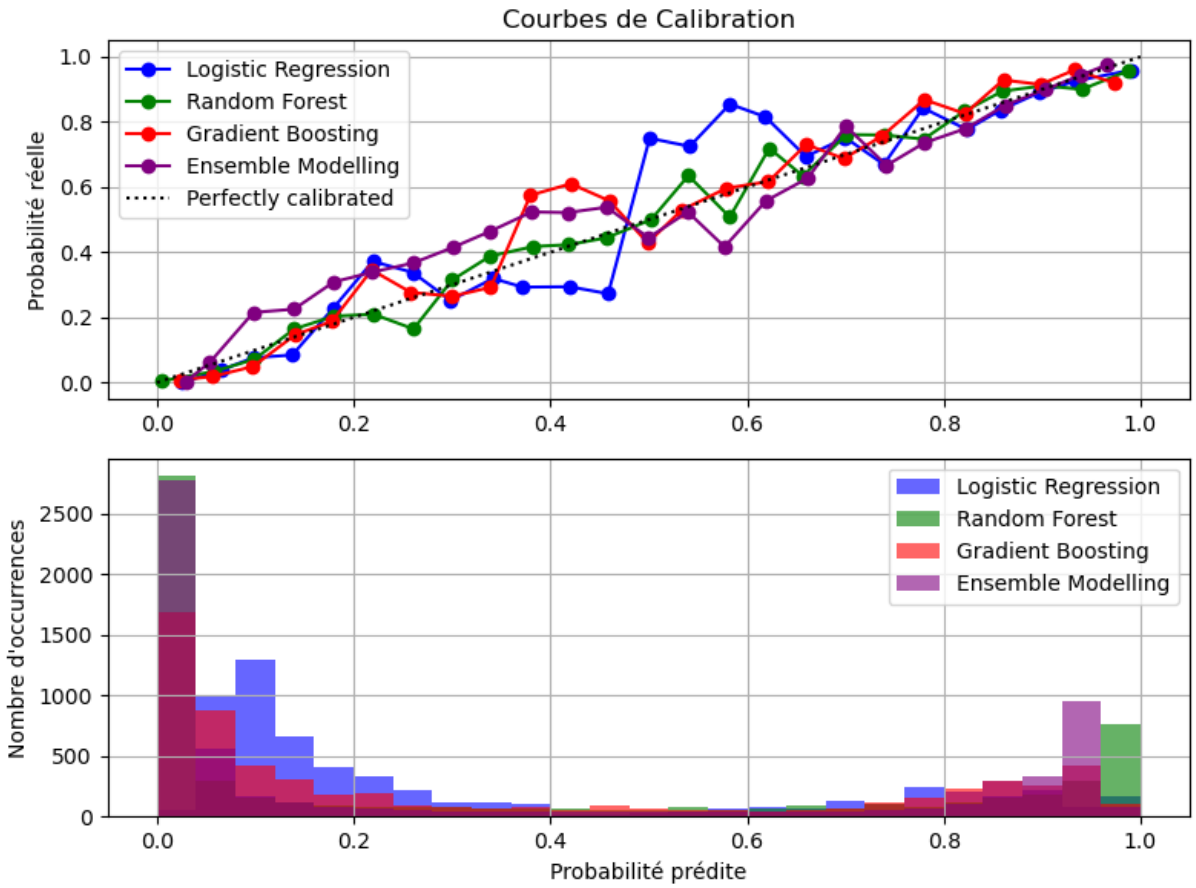


Figure 10 - Courbes de Calibration des probabilités

En analysant ce graphique, on peut constater que les modèles ont globalement une bonne calibration car elles suivent de près la ligne de calibration parfaite. Cependant il y a des fluctuations, surtout aux extrêmes des probabilités (0 et 1), indiquant une certaine incertitude ou un manque de calibration dans ces zones. Le modèle de Gradient Boosting semble avoir des prévisions légèrement surévaluées ou sous-évaluées par rapport aux autres modèles. De plus en analysant l'histogramme des probabilités prédites, on observe une grande concentration des prédictions autour des probabilités de 0 et 1 pour certains modèles, notamment pour le modèle d'ensemble et Random Forest. Cela peut indiquer que ces modèles sont très sûrs de leurs prédictions (ils prédisent souvent des probabilités proches de 0 ou 1). Les autres modèles montrent une distribution plus étalée des probabilités prédites, ce qui peut refléter une grande incertitude ou ne prédiction plus modérée. Cette distribution étendue pourrait également indiquer qu'ils ont une approche plus conservatrice en évitant les prédictions extrêmes à moins d'être très confiants. En conclusion, le modèle d'ensemble est le meilleur choix pour la détection des entreprises innovantes comparé autres modèles. Ce modèle montre une bonne calibration et une performance élevée sur les métriques de classification ainsi que sur l'analyse de la courbe

ROC (annexe 5). En combinant plusieurs modèles, cette méthode équilibre les forces et faiblesses de chacun, produisant ainsi les meilleures prédictions de probabilités

7.7 Entreprises identifiées

SIREN	NOMEN	DEP	Commune
319632790	arkema france	04	Château-Arnoux-Saint-Auban
388216616	anticor chimie	04	Peyruis
792068272	a2i didact	04	Manosque
518330634	mirca	04	Manosque
414520403	alpes ingenierie informatique	04	Oraison
788991651	bmc	05	Chorges
387050115	acacias ctre maladies ..	05	Briançon
508289931	hydretudes alpes du sud	05	Gap
491142337	instore solution	05	Gap
824429344	mcs	05	Gap
348099987	nge fondations	06	Drap
380375097	cogedim gestion	06	Nice
838877207	himydata	06	Valbonne
325089589	airbus ds geo sa	06	Valbonne
415550284	v. mane fils	06	Le Bar-sur-Loup
440571784	gecko software	13	Aix-en-Provence
378159818	holding bernard blachere	13	Châteaurenard
489721704	ip energy	13	Gardanne
508973161	webikeo	13	Aix-en-Provence
407710318	cpp digital media	13	Rognes
844787234	cnim environnement & energie epc	83	La Seyne-sur-Mer
440327518	jcr	83	Fréjus
518127667	marcel	83	Ramatuelle
804820082	interact software	83	Toulon
507399236	elap	83	Brignoles
312670730	sormaf	84	Cavaillon
456500537	dalkia	84	Avignon

392828083	eutrope	84	L'Isle-sur-la-Sorgue
809024516	sra sud est	84	Velleron
512008418	ajr conseil	84	Avignon

TABLE 5 - Top 5 des entreprises identifiées par département

Nous avons retenu le modèle d'ensemble pour la détection des entreprises qui pourrait être des lauréats potentiels pour le Fonds France 2030. Ce modèle a permis d'identifier plusieurs entreprises dans les six départements de la région, illustrant une diversité d'industries et de secteurs innovants. Le tableau ci-dessus montre le Top5 des entreprises identifiées par département. A titre illustratif, dans le département des Alpes-de-Haute-Provence (04), on remarque Arkema France, un acteur majeur de l'industrie chimique en France, qui peut être un lauréat potentiel à France 2030. Dans le département des Alpes-Maritimes (06) nous avons Airbus DS Geo SA qui est impliqué dans les solutions géospatiales et les services de surveillance terrestre. Dans le département des Bouches-du-Rhône (13) nous avons Weibikeo qui propose des solutions webinaires et de communication digitale. Dans le département du Var (83) nous avons CNIM Environnement & Energie EPC qui est spécialisé dans les technologies vertes et les solutions énergétiques durables. Dans le département des Hautes-Alpes (05) nous avons Hydretudes Alpes du Sud qui est spécialisé dans les études et la gestion des ressources hydrauliques. Ces entreprises détectées par le modèle montrent un potentiel élevé pour répondre aux critères du Fonds France 2030. Sur la base de ces résultats, la DREETS Provence-Alpes-Côte d'Azur mènera une analyse plus approfondie de ces entreprises pour éventuellement leur proposer un accompagnement personnalisé dans le cadre de France 2030.

8 Discussions et Recommandations

Nous avons utilisé quatre modèles d'apprentissage automatique dans cette étude pour pouvoir répondre à la question de la détection des entreprises innovantes en PACA. Chacun de ces modèles présentent ces avantages et ses limites. La régression logistique, simple et interprétable, a montré des performances correctes mais limitées par rapport aux autres modèles plus sophistiqués, de même que sa courbe de calibration des probabilités. La Forêt aléatoire offre une précision et un rappel élevés, réduisant les risques de surajustement grâce à son approche d'ensemble. Le Boosting de Gradient, efficace pour capturer les relations non linéaires. Ce modèle a montré une bonne performance globale mais des variations dans la calibration des probabilités, indiquant un besoin d'optimisation supplémentaire. Enfin, le modèle d'ensemble,

combinant les forces des trois autres modèles, a montré les meilleures performances globales, offrant un bon équilibre entre la précision et rappel et une calibration des probabilités satisfaisante. Par ailleurs, notons que pouvons avons des résultats meilleurs et plus satisfaisante avec ces modèles, si d'une part nous utilisons des techniques de calibrage comme la régression isotonique ou la régression logistique platt pour améliorer la précision des probabilités prédictives. D'autre part, pour améliorer les performances des modèles de prédiction, on peut enrichir les données avec d'autres variables supplémentaires. Voici quelques recommandations sur les variables à considérer :

- Les variables Financières liés aux revenus annuels, aux bénéfices nets, aux flux de trésorerie opérationnels, aux flux de trésorerie d'investissement et de financement, au ratio d'endettement, et à la capacité de remboursement ;
- Variables de Performance Opérationnelle comme les dépenses en recherche et développement (R&D), le nombre de brevets, etc.

9 Conclusion

A l'issue de mon alternance au sein du SGAR, j'ai approfondi et amélioré mes acquis théoriques et techniques. Cette année d'apprentissage m'a permis de mettre en pratique mes connaissances à travers diverses tâches et projets. L'alternance m'a permis de faire un pont entre ma formation et les missions de Data analyst au SGAR. Au cours de cette expérience enrichissante, j'ai appris sur mon métier mais j'ai développé des compétences en autonomie, en anticipation de besoins et en résolution de problème. En dehors de mes missions quotidiennes qui sont entre autres l'extraction, le nettoyage, la visualisation et l'analyse, j'ai travaillé à mettre en place un modèle de détection des entreprises innovantes en PACA à partir des données de l'entreprise, dont les résultats serviront dans le cadre du pilotage du Fonds France 2030.

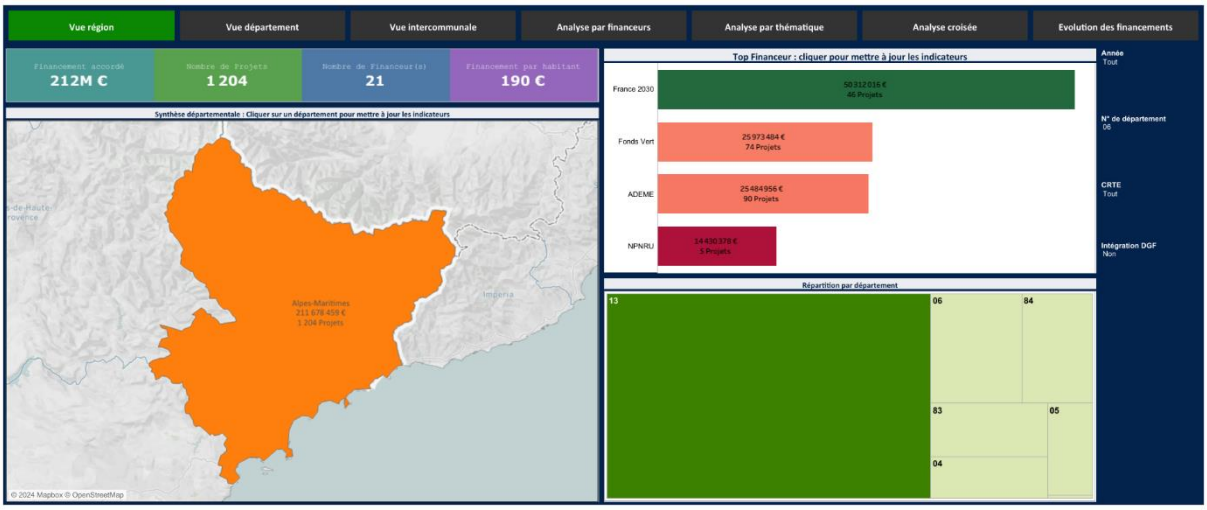
Ce projet avait pour objectif de développer et d'évaluer divers modèles d'apprentissage automatique pour identifier des entreprises prometteuses pouvant bénéficier du Fonds France 2030. Nous avons comparé plusieurs modèles, incluant la régression logistique, les forêts aléatoires, le boosting de gradient, ainsi qu'un modèle d'ensemble combinant ces trois approches. Les principaux résultats suggèrent que le modèle d'ensemble est plus performant, offrant une précision et une robustesse supérieures par rapport aux autres modèles individuels. Cela a été confirmé par les métriques de classification (précision, rappel, F1-score) et par les courbes de calibration. La diversité et la qualité des variables utilisées se sont avérées cruciales pour la performance des modèles. L'intégration de nouvelles variables, notamment financières, de marché, opérationnelles, est recommandée pour améliorer encore les prédictions futures. Pour les travaux futurs, il serait bénéfique de continuer à expérimenter avec différents modèles et techniques d'ensemble pour améliorer la précision et la robustesse des prédictions, d'élargir les données en intégrant des données plus diversifiées, notamment des indicateurs de performance environnementale et sociale, développer des systèmes automatisés pour la collecte de données et l'évaluation continue des entreprises afin de maintenir à jour les modèles prédictifs. Enfin, ce projet montre que l'application des modèles d'apprentissage automatique peut considérablement améliorer l'identification en avance de phase des bénéficiaires pour des initiatives de financement telles que le fonds France 2030. Les résultats obtenus et les recommandations formulées offrent une feuille de route pour les améliorations futures, garantissant une utilisation optimale des ressources et un soutien accru aux entreprises innovantes.

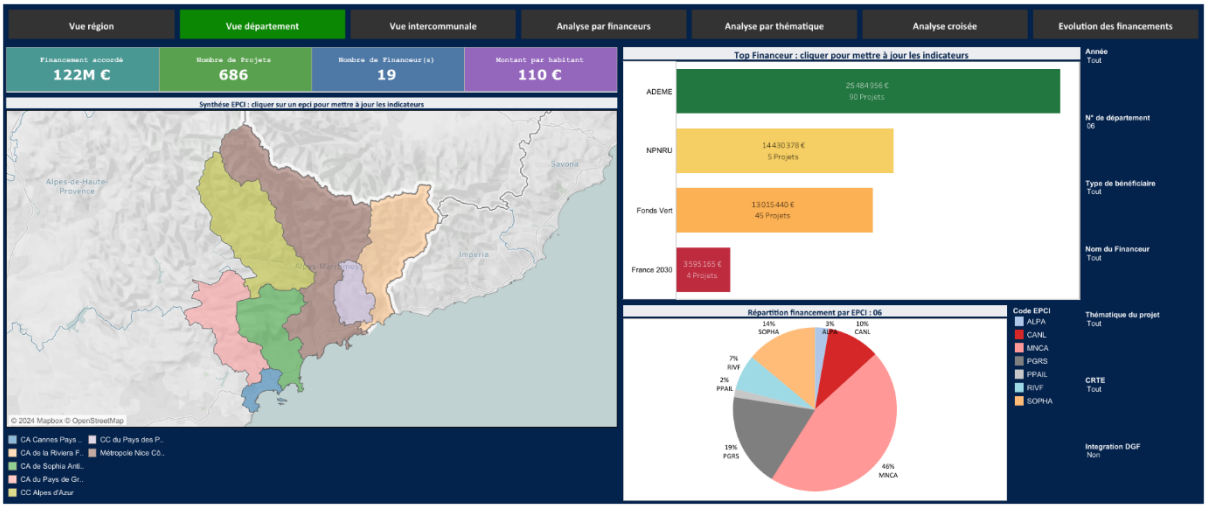
10 Références Bibliographiques

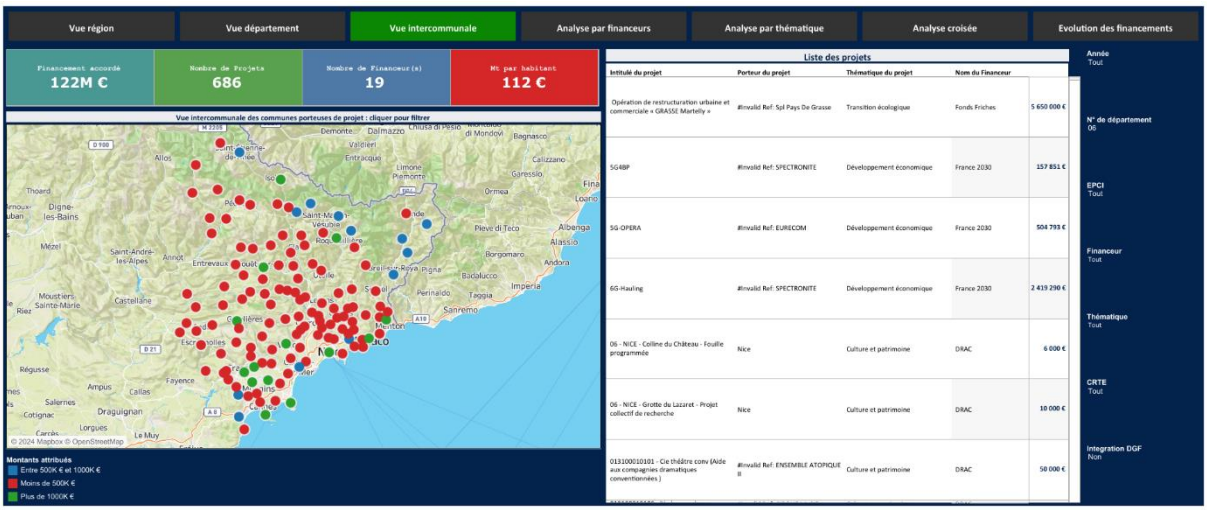
1. OCDE. (2018). Manuel d'Oslo : Directives pour le recueil et l'interprétation des données sur l'innovation. Organisation de Coopération et de Développement Économiques.
2. Nguyen, T., et al. (2020). Predicting SME Innovation with Machine Learning: A Financial and Non-Financial Data Approach. Journal of Business Research.
3. Yu, X., et al. (2019). Clustering Innovative Firms in the Chinese Technology Sector: A Machine Learning Approach. Technological Forecasting and Social Change.
4. Zhang, W., et al. (2018). Using Machine Learning to Predict Firm Innovation from Textual and Financial Data. International Journal of Information Management.
5. Li, M., et al. (2017). NLP-based Analysis of Corporate Publications for Innovation Detection. IEEE Transactions on Knowledge and Data Engineering.
6. Chen, Q., et al. (2018). Adaptive Machine Learning Models for Innovation Trend Tracking. Journal of Technology Transfer.
7. Audretsch, D. B., & Keilbach, M. (2004). Entrepreneurship and Regional Growth: An Evolutionary Interpretation. Journal of Evolutionary Economics.
8. Czarnitzki, D., & Delanote, J. (2013). Young Innovative Companies: The New High-Growth Firms. Industrial and Corporate Change.
9. France 2030 : un plan d'investissement pour la France, https://www.economie.gouv.fr/france-2030
10. Prefecture BDR SGAR, https://www.prefectures-regions.gouv.fr/provence-alpes-cote-dazur/irecontenu/telechargement/113121/846351/file/Prefecture%20BDR%20SGAR%20depliant%20document%20final.pdf
11. SMOTE et données mixtes, traiter les variables catégorielles avec SMOTE-NC, https://kobia.fr/imbalanced-data-smote-nc
12. Déclarer la guerre aux données déséquilibrées : SMOTE, https://www.neosoft.fr/nos-publications/blog-tech/techniques-augmentation-dataset-smote/
13. Le SGAR et ses missions, https://www.prefectures-regions.gouv.fr/provence-alpes-cote-dazur/provence-alpes-cote-dazur/Region-et-institutions/Organisation-administrative-de-la-region/Le-SGAR/Le-SGAR-et-ses-missions

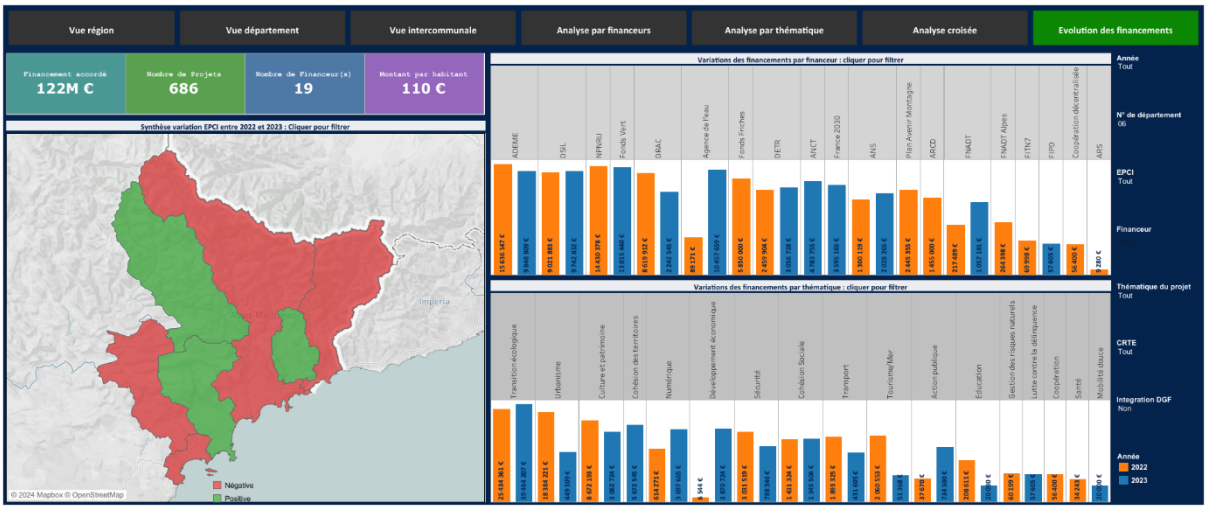
11 Annexes

1. Dashboard des données financières

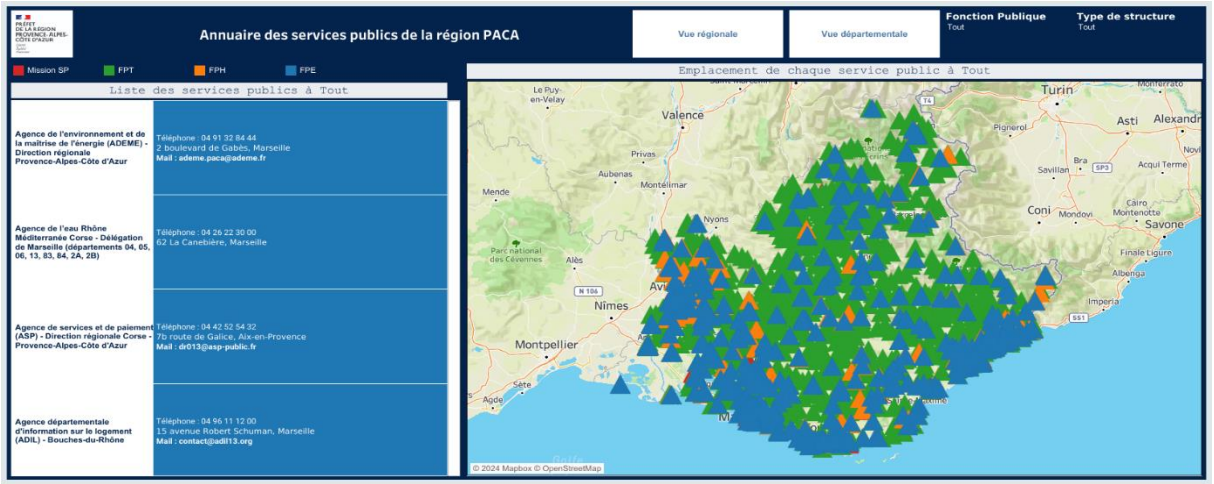


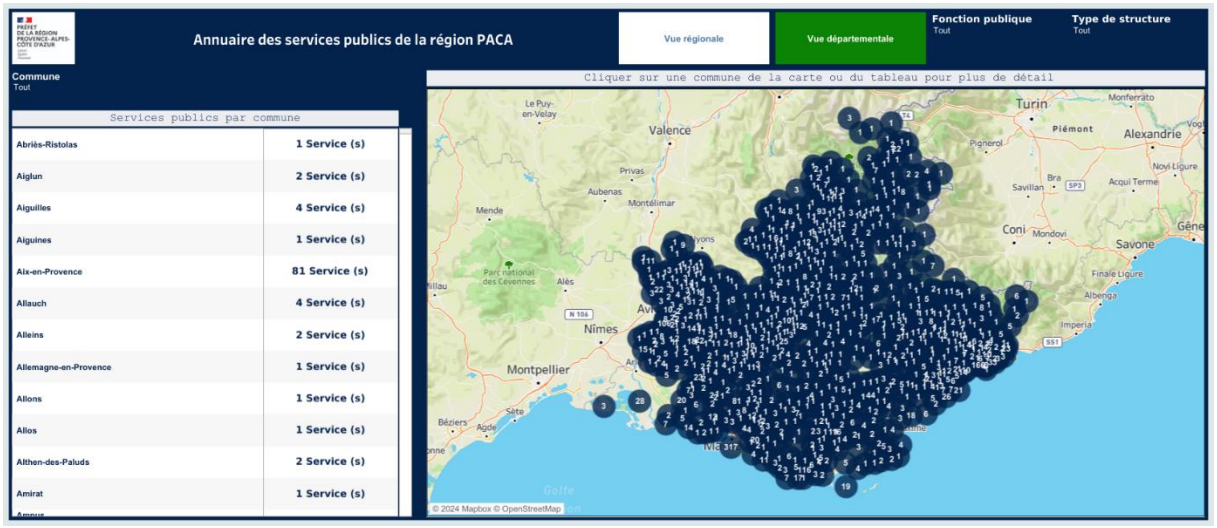


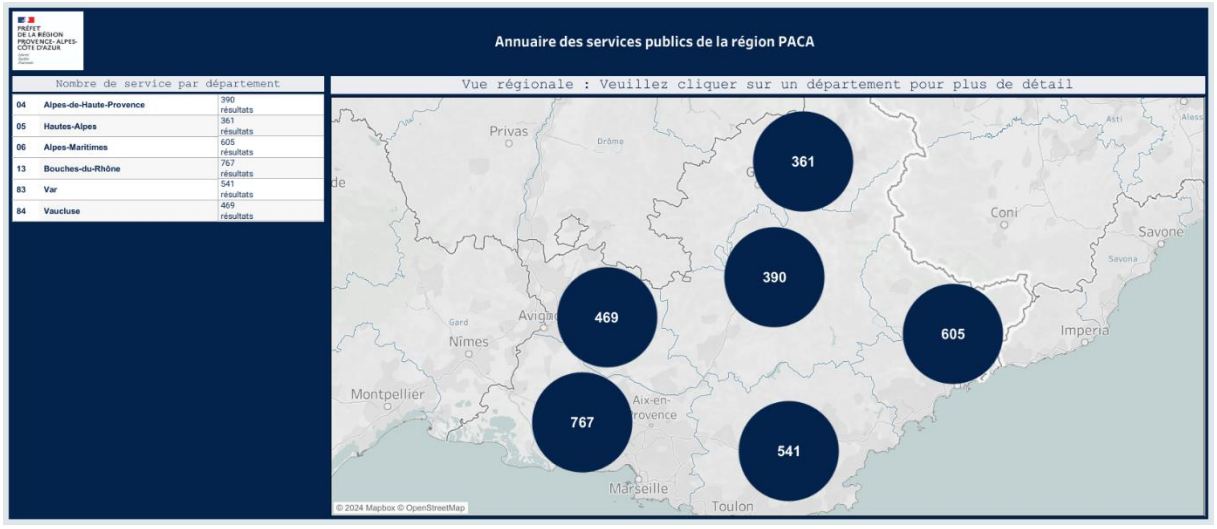




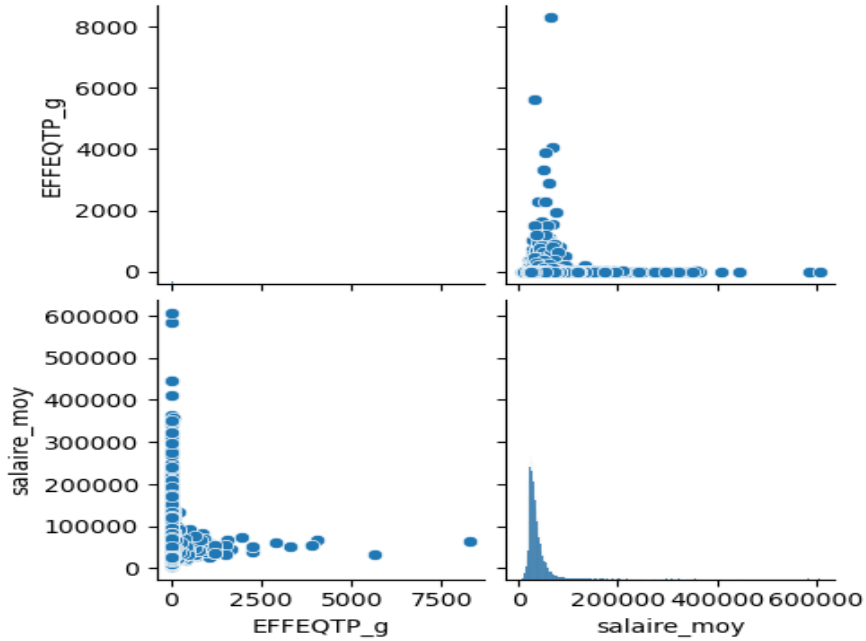
2. Dashboard données des services publics



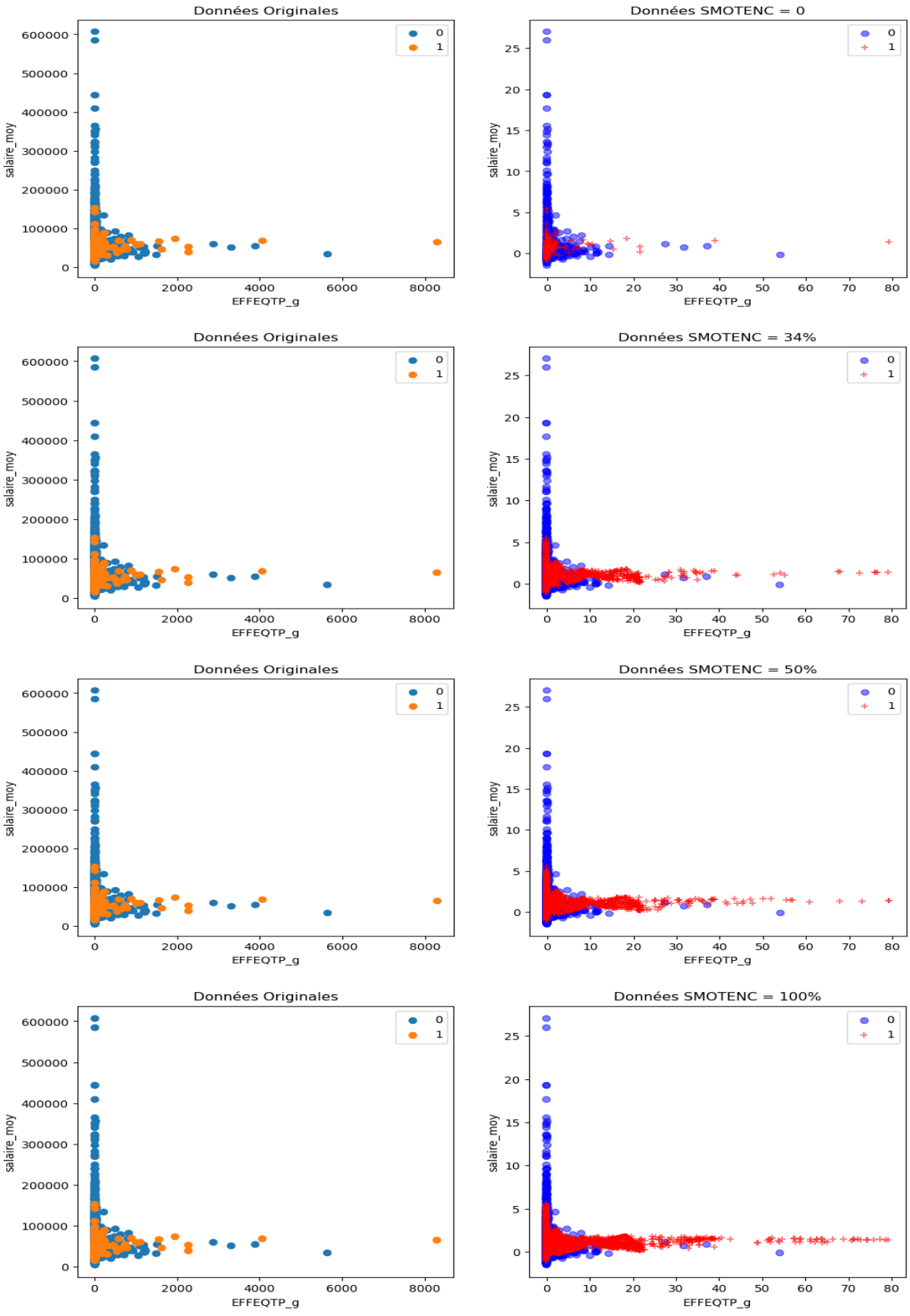




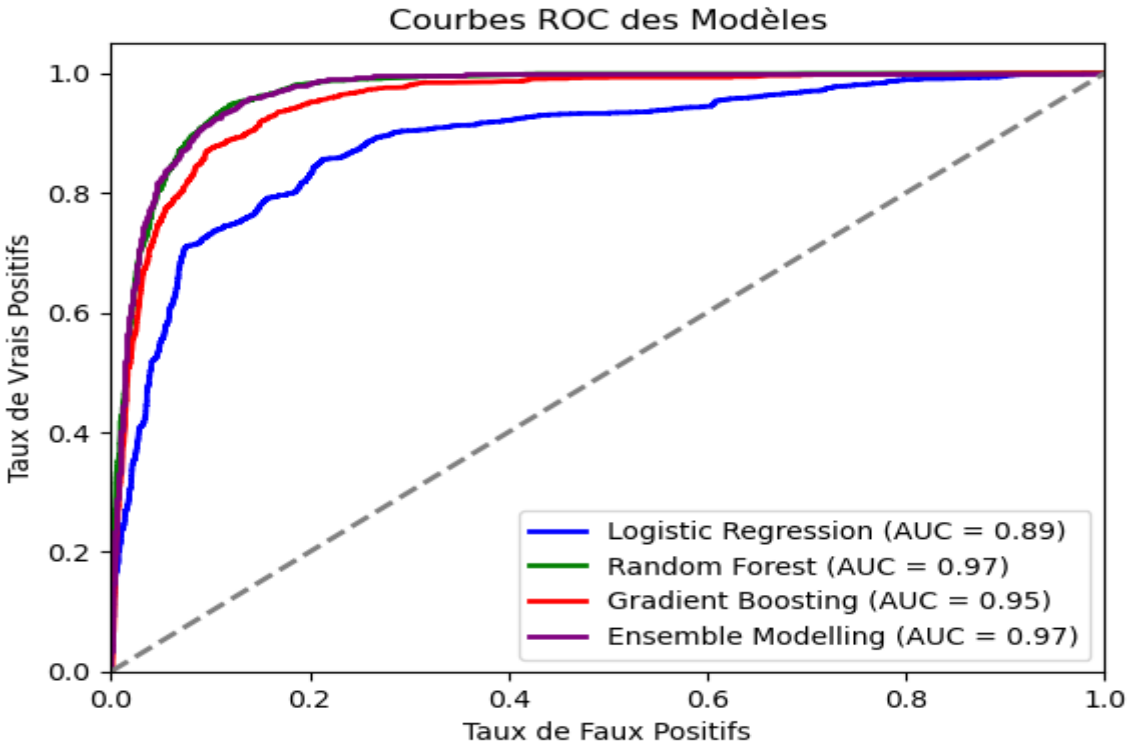
3. Présentation de la distribution des variables quantitatives



4. Effets de SMOTENC sur les données



5. Courbe ROC


Courbes ROC des Modèles
Taux de Vrais Positifs
Taux de Faux Positifs
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.89)
Random Forest (AUC = 0.97)
Gradient Boosting (AUC = 0.95)
Ensemble Modelling (AUC = 0.97)